VLBiMan++: Expanding the Generalization Boundary of Vision-Language Anchored One-Shot Bimanual Manipulation

Huayi Zhou, *Member, IEEE,* Wei Gao, Yiyang Han, Kui Jia, *Member, IEEE,* Hui Huang, *Senior Member, IEEE*

***Abstract*—Generalizable bimanual robotic manipulation requires a reusable task prior that can persist across increasingly diverse tasks, objects, scenes, embodiments, and execution conditions, thus avoiding the prohibitive cost of large-scale teleoperated demonstrations and policy retraining. In this work, we present VLBiMan++, an extended framework that expands the generalization boundary of vision-language anchored one-shot bimanual manipulation. Starting from a single human demonstration, VLBiMan++ performs task-aware decomposition to identify reusable and adaptable skill components, and employs vision-language grounded geometric adaptation to transfer these skills to novel configurations without retraining. Building on this foundation, we systematically extend generalization along five dimensions: task generalization through diverse and long-horizon skill compositions; object generalization across unseen categories, varying geometries, and more complex articulated or deformable objects; scene generalization under clutter, occlusion, and dynamic interference; embodiment generalization across heterogeneous dual-arm robotic platforms; and deployment generalization through prolonged closed-loop execution under repeated external perturbations. To support this broader scope, we further introduce object-state-aware adaptation and lightweight trajectory optimization mechanisms that accommodate changes beyond simple rigid-body 6-DoF pose variations while preserving reliable bimanual coordination. Extensive real-world experiments demonstrate that VLBiMan++ maintains strong task success and adaptation capability across these increasingly challenging settings. Overall, VLBiMan++ advances one-shot bimanual manipulation from demonstrating isolated transferability toward a more systematic and scalable framework for generalization across tasks, objects, scenes, embodiments, and long-term deployment conditions. The project link is https://hnuzhy.github.io/projects/VLBiManPlus.**

***Index Terms*—Bimanual Manipulation, One-Shot Demonstration Learning, Vision-Language Grounding, Skill Generalization.**

I. Introduction

RECENT years have witnessed remarkable progress in embodied robotic manipulation, particularly through visuomotor imitation learning based on large-scale teleoperated demonstrations [1]–[4]. By collecting large quantities of real-world demonstrations, recent Vision-Language-Action (VLA) models [5]–[7] learn to directly map multimodal observations to robot actions, implicitly encoding task-, object-, and environment-specific complexities into high-dimensional latent representations. This paradigm has demonstrated impressive capabilities on challenging high-degree-of-freedom manipulation problems, including bimanual and long-horizon tasks, as evidenced by the ALOHA family [8]–[11], RDT-1B [12], π_0 [13], and FAST [14]. However, its scalability remains fundamentally constrained by the cost of data collection and policy retraining: adapting an existing policy to new tasks, object configurations, or robotic embodiments often requires substantial demonstration re-collection and retraining. Such a paradigm is therefore difficult to scale to the unstructured combinations of tasks, objects, scenes, and robot platforms encountered in real-world deployment.

An alternative direction is to exploit pretrained foundation models and structured intermediate representations to decouple semantic understanding from low-level control. Recent modular robotic systems leverage Large Language Models (LLMs) [15] and Vision-Language Models (VLMs) [16], [17] for task interpretation, semantic grounding, and object-centric perception, while delegating motion generation to rule-based controllers, learned visuomotor policies, or reusable skill libraries [18]–[20]. Besides, reinforcement learning in simulation also serves as a strategy for learning skill-specific controllers [21]–[23]. Another particularly effective strategy is to introduce compact and transferable geometric representations, such as keypoints, affordances, and object-centric correspondences, as a bridge between perception and action. For example, ReKep [24] constrains trajectory optimization using predicted relation points, MOKA [25] identifies functional regions through multimodal reasoning, and RobotPoint [26] extracts task-relevant point clusters from visual observations. These studies demonstrate that object-centric abstractions can substantially improve manipulation transferability across changing viewpoints and instances. Nevertheless, existing approaches typically focus on a limited aspect of generalization and often depend on either extensive task-specific data, specialized perception models, or planning pipelines whose execution reliability can deteriorate as the manipulation scenario becomes increasingly diverse.

Our previous conference work, **VLBiMan** [27], explored a complementary hypothesis: rather than repeatedly relearning

H. Zhou and H. Huang are with Guangdong Provincial Key Laboratory of Visual Media and Multidimensional Intelligence, College of Computer Science and Software Engineering, Shenzhen University. K. Jia is with School of Data Science, The Chinese University of Hong Kong, Shenzhen. W. Gao and K. Jia are with DexForce, Shenzhen. Y. Han is with the Faculty of Engineering, Department of Computing, Imperial College London

This work was supported by the Guangdong Provincial Key Field R&D Program (Project No. 20240104), and was also funded by the Shenzhen Science and Technology Major Project (No. 202402002 and ZDCT20250901113000001).

Manuscript received April 25, 2026; revised August 25, 2026.

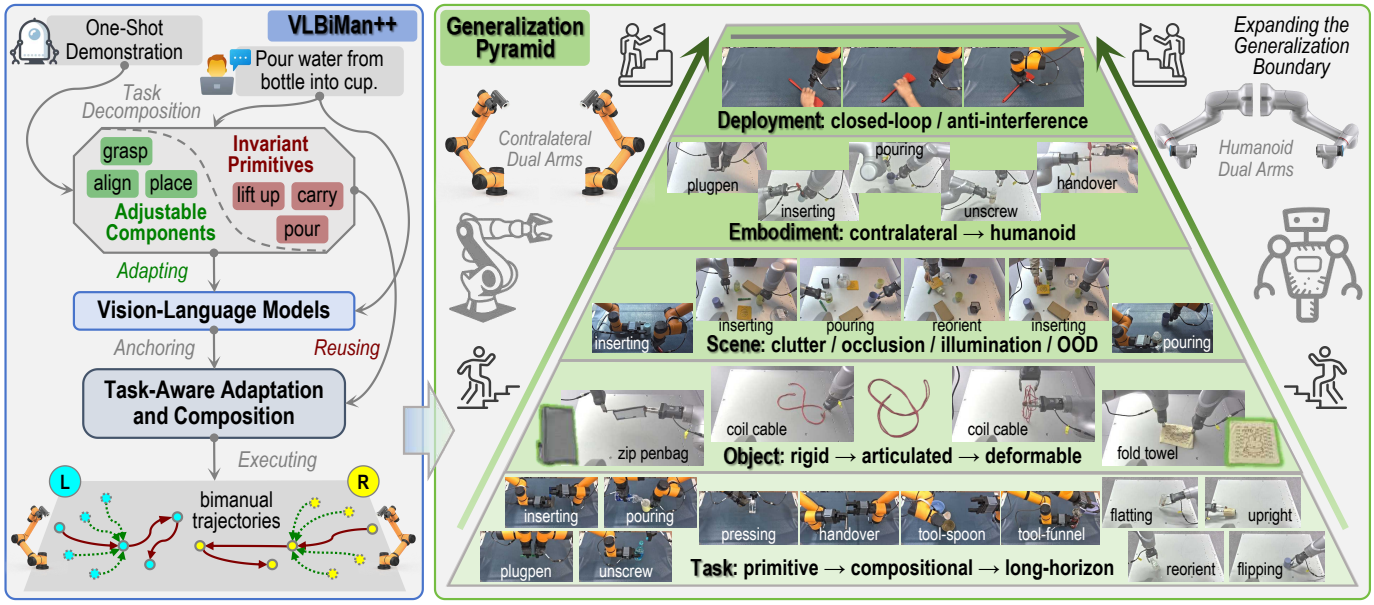


Fig. 1. *Left*: Taking pouring water as an example, we sketch the entire process of VLBiMan++ based on the one-shot demonstration. *Right*: Diagram illustration of the proposed **Generalization Pyramid**: **Task** → **Object** → **Scene** → **Embodiment** → **Deployment**. VLBiMan++ aims to investigate how far a single human demonstration can serve as a reusable manipulation prior across increasingly diverse tasks, objects, scenes, embodiments, and execution conditions.

or replanning an entire manipulation policy, a single human demonstration can serve as a structured prior from which the essential and reusable parts of a task are extracted. VLBiMan performs (1) *Task-Aware Bimanual Decomposition*, (2) *Vision-Language Anchored Adaptation*, and (3) *Autonomous Trajectory Composition* to separate invariant skill components from scene-dependent variations, align the latter through VLM-grounded object representations, and synthesize executable trajectories without policy retraining. The central design principle is that ***what to achieve matters more than how to execute it***: task-relevant spatial and functional relationships should be preserved, whereas incidental absolute poses and trajectories can be adapted. This formulation enables effective transfer from a single demonstration across object placements, category-level instances, long-horizon skill compositions, and heterogeneous robotic embodiments. However, the original VLBiMan [27] primarily establishes the feasibility of this paradigm on a controlled collection of rigid-object manipulation tasks. A more fundamental question therefore remains: *how far can a one-shot human demonstration generalize when the task, object, scene, embodiment, and execution conditions progressively depart from those observed in the demonstration?*

In this journal extension, we answer above-mentioned question by introducing **VLBiMan++**, which expands the generalization boundary of vision-language anchored one-shot bimanual robotic manipulation along a systematic **Generalization Pyramid**: *Task Generalization* → *Object Generalization* → *Scene Generalization* → *Embodiment Generalization* → *Deployment Generalization*. At the ***task*** level, we substantially broaden the experimental suite to cover more diverse bimanual behaviors, multi-stage interactions, and long-horizon skill compositions, probing whether atomic skills extracted from a single demonstration can support broader task structures. At the ***object*** level, we move beyond seen and category-level novel rigid objects to more challenging unseen object types and investigate articulated and deformable objects, extending the original adaptation formulation toward changes in object configuration and physical state rather than simple rigid-body pose variation. At the ***scene*** level, we evaluate VLBiMan++ under cluttered environments containing semantic distractors, partial occlusions, and spatial interference, testing whether vision-language anchoring can continuously identify and track the task-relevant objects despite substantial perceptual ambiguity. At the ***embodiment*** level, we further validate transfer across heterogeneous dual-arm platforms, demonstrating that the task prior extracted from a human demonstration can remain effective under substantially different kinematic and hardware configurations. Finally, at the deployment level, we conduct prolonged closed-loop execution with repeated object perturbations and external disturbances, examining not only instantaneous task success but also the stability and persistence of adaptation over extended operation.

To sum up, VLBiMan++ further extends the original framework to support broader object-state variations and deployment conditions while preserving its training-free, task-grounded philosophy. Beyond adapting rigid-body position and orientation, the enhanced object-centric adaptation accommodates articulated configurations and deformation, while the modular *perception–adaptation–composition* pipeline enables repeated re-observation and re-alignment under object displacement and environmental disturbances. Accordingly, VLBiMan++ treats a human demonstration as a reusable task prior and systematically expands its generalization boundary from task composition and novel object states to cluttered scenes, heterogeneous embodiments, and prolonged closed-loop execution, providing a more comprehensive pathway toward scalable bimanual manipulation that reuses rather than repeatedly relearns human demonstrations. Our contributions are as follows:

- We introduce VLBiMan++, a substantial extension of our previous work VLBiMan, and formulate a Generalization

Pyramid for one-shot bimanual manipulation that systematically studies generalization across tasks, objects, scenes, embodiments, and long-term deployment conditions.
- We extend vision-language anchored adaptation beyond rigid-object pose transfer toward more complex object-state variations, including articulated and deformable manipulation, enabling the reusable prior to remain effective under substantially broader geometric and physical changes.
- We establish a substantially expanded real-world evaluation suite covering unseen object categories, cluttered and dynamically perturbed scenes, new heterogeneous dual-arm embodiments, and prolonged closed-loop execution, providing comprehensive evidence of VLBiMan++'s robustness, versatility, and deployment stability.
- We provide extensive analyses of the expanded generalization boundary, including failure modes, robustness under persistent disturbances, and cross-condition transferability, revealing both the practical strengths and remaining limitations of one-shot bimanual manipulation.

II. Related Works

Generalizable Representations for Manipulation. Traditional robotic manipulation often relied on structured representations built upon strong priors [28]–[31], such as object geometry or rigid-body assumptions, typically through 6D pose estimation or manually specified grasp configurations, which are difficult to scale to unstructured environments. With the rise of data-driven techniques, more flexible representations have emerged, including keypoints [32]–[35], affordances [36]–[39], dynamic flow fields [40], [41], and invariant object-centric correspondences [42]–[44]. More recent works further extend these representations to articulated or deformable objects and leverage human demonstrations to retarget 3D hand trajectories to robots [45]–[48]. However, such approaches often require private datasets, retraining, object-specific models, or complex retargeting pipelines, limiting their scalability across diverse object states. In contrast, VLBiMan++ employs lightweight object-centric representing points together with vision-language grounding and extends adaptation beyond rigid pose changes toward more diverse geometric, articulated, and deformable object variations, without retraining.

Efficient Bimanual Robotic Manipulation. Recent advances in bimanual manipulation have demonstrated the capability of large Vision-Language-Action (VLA) models [12]–[14] trained on extensive teleoperated demonstrations [1]–[4]. However, scaling these systems to new tasks, objects, or embodiments often requires substantial additional data collection and retraining. Alternative efforts exploit large-scale Internet [49]–[51] or egocentric human-hand videos [52]–[56], yet the human-robot embodiment gap limits direct transfer. Other methods learn visuomotor policies [18], [19] from relatively small real-world datasets, but their generalization remains constrained by task-specific training. One-shot imitation learning [57]–[65] substantially reduces data requirements, while the high-dimensional and coordinated nature of bimanual manipulation remains challenging. In contrast, VLBiMan++ reuses task-invariant atomic skills extracted from a single bimanual demonstration and adaptively composes them across tasks, object states, scenes, and embodiments, providing a training-free route toward broader and more scalable generalization.

Large Foundation Models for Robotics. Integrating LLMs and VLMs into robotics has become a prominent direction for building generalizable embodied agents [66]–[69]. LLMs are commonly used for high-level task understanding and planning, such as decomposing instructions into executable subtasks or generating programs [70]–[73], while VLMs provide semantically grounded object detection and segmentation, often complemented by Visual Foundation Models (VFMs) [74], [75] for keypoints [32]–[34] and dense correspondences [42], [43]. Recent systems such as ReKep [24], MOKA [25], RobotPoint [26], and RAM [76] combine foundation models within modular *perceive-understand-plan-act* pipelines for zero-shot manipulation. Despite their strong generalization, these methods often rely on extensive prompt engineering and ambiguous intermediate representations that require additional post-processing or planning. VLBiMan++ instead grounds manipulation adaptation in a precisely labeled one-shot demonstration, using VLMs to identify task-relevant objects and lightweight geometric anchors to selectively adapt only the components affected by object, scene, or embodiment changes. This design further extends foundation-model-guided manipulation toward articulated and deformable objects, cluttered environments, and prolonged closed-loop deployment without policy retraining.

Several concurrent studies have recently explored learning manipulation skills from a single human demonstration, sharing VLBiMan++'s goal of reducing demonstration and retraining costs. Multi-Task Trajectory Transfer (MT3) [77] decomposes trajectories into sequential alignment and interaction phases and employs retrieval-based generalization, enabling efficient transfer across a large collection of tasks. Behavior Prompting Policy (BPP) [78] formulates demonstrations as behavior prompts and learns an in-context visuomotor policy that maps a demonstration and current observation directly to robot actions, while HOST [79] acquires skills from a single human video through self-grounded prediction and a shared task-progress manifold for embodiment adaptation. While these works demonstrate the strong potential of one-shot skill acquisition, VLBiMan++ differs in its emphasis on training-free, explicitly task-aware decomposition and vision-language anchored adaptation for bimanual manipulation, where invariant skill components are selectively reused and variable object-state components are geometrically adapted rather than inferred by a learned visuomotor policy.

III. Methodology

This section presents the complete pipeline of **VLBiMan++** (refer Fig. 2), extending vision-language anchored one-shot manipulation toward broader generalization across tasks, object states, scenes, and long-term deployment. We first formalize the problem and generalized manipulation states in Sec. III-A, and then introduce three interconnected components: (1) *Task-Aware Bimanual Decomposition* in Sec. III-B, which decomposes a single demonstration into reusable atomic

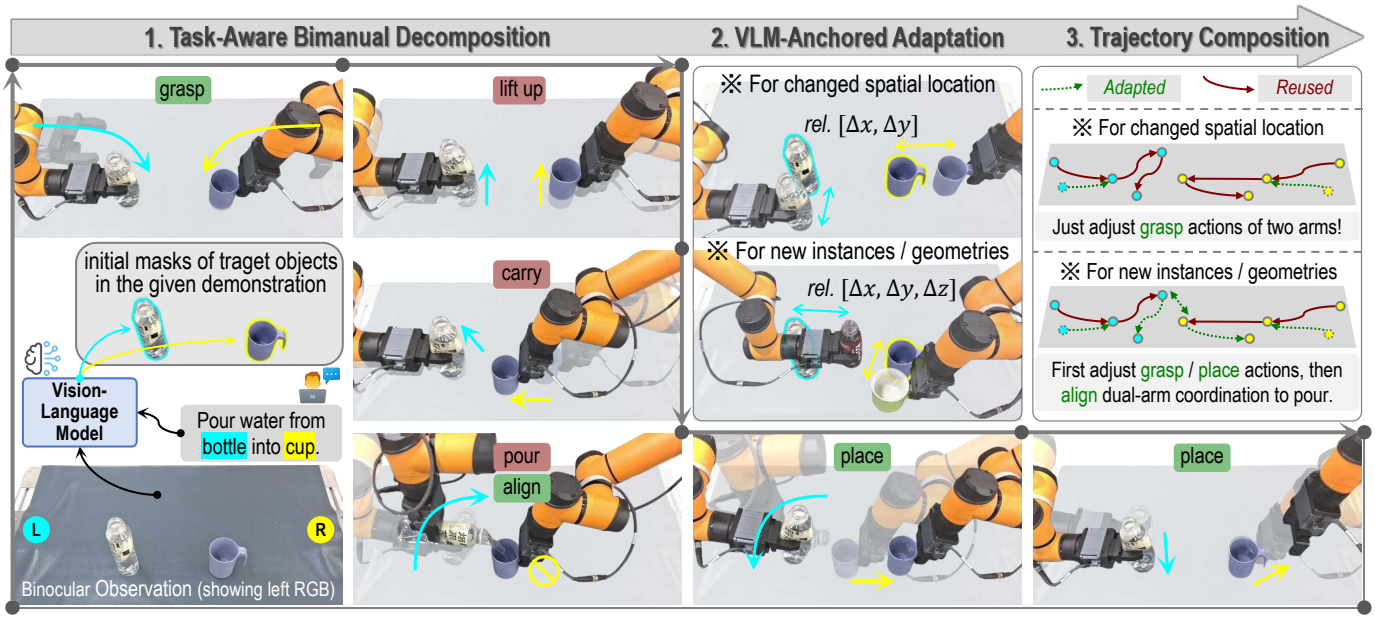


Fig. 2. Framework of the expanded **V**ision-**L**anguage Anchored **Bi**manual **Mani**pulation (**VLBiMan++**). Taking the pouring water as an example, the paradigm consists of three stages (*e.g.*, decomposition, adaptation, and composition) based on a given demonstration. VLBiMan++ can achieve generalization of unseen spatial placements, category-level new instances, non-rigid object state variations, and heterogeneous novel embodiments under the same task.

skills and characterizes their state dependence; (2) *Vision-Language Anchored Adaptation* in Sec. III-C, which grounds task-relevant objects and adapts skills to rigid, articulated, and deformable object states; and (3) *Autonomous Trajectory Composition* in Sec. III-D, which synthesizes executable bimanual trajectories through kinematic refinement, collision compensation, and closed-loop re-observation.

A. Preliminaries

Given a concise textual description of a bimanual manipulation task and a single human demonstration collected in a canonical scene, our goal is to synthesize executable dual-arm trajectories for unseen conditions without policy retraining. Let $\mathcal{T}$ denote the task description and $\mathcal{D} = \{(\mathcal{O}_t, \mathcal{A}_t)\}_{t=1}^{T}$ denote the one-shot demonstration, where $\mathcal{O}_t$ is the multi-modal observation at timestep t and $\mathcal{A}_t$ is the corresponding bimanual action (*e.g.*, 6-DoF end-effector poses of both arms, and gripper states). Given a novel scene state $\mathcal{S}_{\text{new}}$, which may involve changes in object identity, geometry, configuration, spatial arrangement, scene clutter, or robot embodiment, VLBiMan++ generates an adapted trajectory:

$$\mathcal{F}_{\text{VLBiMan++}} : (\mathcal{T}, \mathcal{D}, \mathcal{S}_{\text{new}}) \mapsto \widetilde{\mathcal{D}}_{\text{new}} = \{\widetilde{\mathcal{A}}_t^{\text{new}}\}_{t=1}^{T'}, \quad (1)$$

Here, $\mathcal{S}_{\text{new}}$ is not restricted to simple object relocation, scene rearrangement or category-level instance variation, but may further involve unseen object categories, articulated configurations, deformable states, cluttered environments, and changes in the underlying robot embodiment. The $\widetilde{\mathcal{A}}_t^{\text{new}}$ denotes the synthesized bimanual trajectory adapted to $\mathcal{S}_{\text{new}}$. Rather than reproducing the demonstration indiscriminately, VLBiMan++ preserves task-relevant invariant structure and selectively adapts state-dependent components. This requires addressing three core challenges: (1) *Task-object semantic grounding*, which associates $\mathcal{T}$ with relevant objects $o_k{}_{k=1}^{K}$ through visual-language grounding; (2) *Executable module decomposition*, which partitions $\mathcal{D}$ into temporally ordered motion primitives $\mathcal{M}_i$ with discrete boundaries t_i and identifies their state dependence; and (3) *Kinematically feasible trajectory composition*, which synthesizes adapted trajectories $\widetilde{\mathcal{A}}_t^{\text{new}}$ under geometric and physical constraints.

Accordingly, VLBiMan++ views a single given demonstration as a reusable task prior whose generalization spans progressively broader variations, from *task* composition and *object* state changes to *scene* conditions, robot *embodiment*, and long-term *deployment*. Its objective is to maximize reuse of demonstration-derived manipulation knowledge while adapting only the components affected by the current scene state $\mathcal{S}_{\text{new}}$, thereby minimizing additional supervision, retraining, and full-trajectory replanning.

B. Task-Aware Bimanual Decomposition

To construct reusable manipulation priors from a single demonstration, VLBiMan++ decomposes the demonstrated bimanual trajectory into temporally coherent segments and characterizes their dependence on object states. The procedure consists of *Spatiotemporal Trajectory Segmentation*, *Atomic Skill Extraction*, and *State-Aware Skill Abstraction*.

1) Spatiotemporal Trajectory Segmentation: We record each one-shot demonstration using a third-person stereo RGB camera at 10 FPS, synchronously collecting the 6-DoF end-effector poses and gripper states of both arms. The resulting sequence is $\mathcal{D}=\{(\mathcal{O}_t, \mathcal{A}_t)\}_{t=1}^{T}$, where $\mathcal{A}_t \in \mathbb{R}^{14}$ contains the two end-effector poses and binary gripper states. We adopt a keypose-driven segmentation strategy inspired by discrete motion representations [80]–[83]. Candidate keyposes are automatically detected from motion and state changes, including velocity or acceleration discontinuities and gripper transitions. Each keypose $\mathbf{w}_i$ defines a temporal slot $\tau_i = [t_i, t_{i+1}]$ and a candidate segment $\mathcal{M}_i = \{\mathcal{A}_t\}_{t\in\tau_i}$. Furthermore, because

automatically detected keyposes may not always yield reliable waypoints for subsequent execution connected via inverse kinematics (IK) [84], [85], we apply lightweight human-in-the-loop refinement to adjust, insert, or remove a small number of waypoints when necessary. The refined segmentation is validated by real-robot rollout to ensure temporal consistency, spatial smoothness, and safe bimanual execution:

$$\pi_{\texttt{seg}} : \mathcal{D} \rightarrow \{\mathcal{M}_i\}_{i=1}^{N}, \tag{2}$$

where N denotes the number of temporally ordered motion segments, ranging from a few to over a dozen.

2) Atomic Skill Extraction: We next assign semantic roles to the segments according to the coupling between manipulated objects and robot end-effectors. Before grasping, the object and end-effector are spatially independent, making the corresponding motion generally sensitive to the current object state. After stable grasping, they form a coupled composite system, and subsequent motions can often be reused as task-level skills. Let $\texttt{bind}(o, r, t) \in \{0, 1\}$ indicate whether object o is coupled with end-effector r at time t. Following the original decomposition in VLBiMan [27], a segment $\mathcal{M}_i$ is regarded as invariant when the object remains coupled and its geometry is sufficiently consistent with the demonstration:

$$\begin{aligned} &\forall t \in \tau_i,\ \texttt{bind}(o_k, r, t) = 1, \text{and} \\ &\qquad \texttt{geometry}(o_k) \approx \texttt{geometry}(o_k^{\texttt{demo}}), \end{aligned} \tag{3}$$

where ϵ_g denotes the corresponding tolerance threshold for geometrically equivalent dimensions. Otherwise, the segment is considered potentially adaptable. The given demonstration is therefore decomposed into:

$$\mathcal{D} \Rightarrow \{\mathcal{M}_i^{\texttt{inv}}\}_{i=1}^{N_{\texttt{inv}}} \cup \{\mathcal{M}_j^{\texttt{var}}\}_{j=1}^{N_{\texttt{var}}}. \tag{4}$$

which are subsequently reused or adapted for novel scenes, tasks and object variations. Some illustrations on the pouring water task can be found in Fig. 2 left.

3) State-Aware Skill Abstraction: VLBiMan++ further generalizes this decomposition by explicitly modeling whether a skill depends on the task-relevant object state s_o. Here, s_o may describe pose and geometry for rigid objects, configuration for articulated objects, or state for deformable objects. We distinguish *state-invariant* skills, whose functional execution remains valid despite changes in s_o, from *state-conditioned* skills, whose execution must be adapted accordingly:

$$\begin{cases} \mathcal{M}_i^{\texttt{SI}} \in \mathcal{S}^{\texttt{SI}}, & \partial \mathcal{M}_i / \partial s_o \approx 0, \\ \mathcal{M}_j^{\texttt{SC}}(s_o) \in \mathcal{S}^{\texttt{SC}}, & \partial \mathcal{M}_j / \partial s_o \neq 0, \end{cases} \tag{5}$$

Here, *state-invariant* skills typically correspond to motions such as lifting, carrying, or transporting an already grasped object, whereas *state-conditioned* skills include motions whose feasibility depends on object position, orientation, geometry, articulation, or deformation. Thus, rigid, articulated, and deformable objects can be handled within a unified abstraction: invariant structure is preserved as a reusable skill, while state-dependent components are exposed for subsequent adaptation. The resulting skill set is $\mathcal{S} = \mathcal{S}^{\texttt{SI}} \cup \mathcal{S}^{\texttt{SC}}$ providing a structured task prior for the next stage.

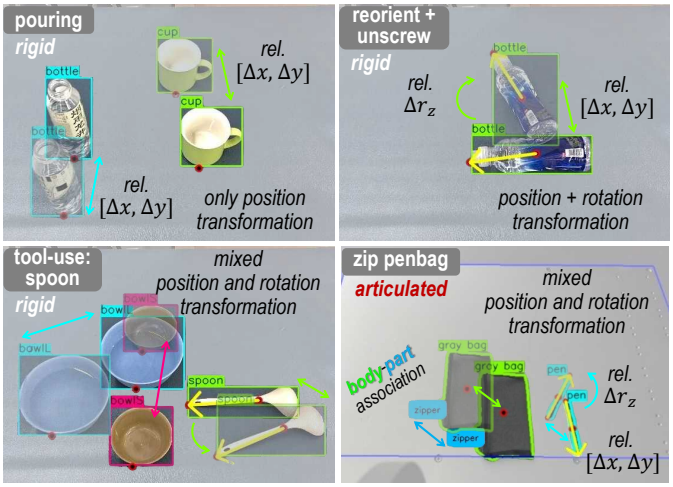


Fig. 3. Illustrations of representative points for target objects in four tasks including `pouring`, `reorient+unscrew`, `tool-use:spoon`, and `zip penbag`. The first three tasks involve only rigid objects, while the last task involves an articulated object. These anchor points will be used to calculate the change in object position and orientation (not always required).

C. Vision-Language Anchored Adaptation

After task-aware decomposition, VLBiMan++ adapts state-conditioned skills to novel object and scene states through vision-language grounding and object-centric geometric reasoning. Rather than estimating a complete physical scene state, we identify task-relevant objects and extract compact anchors that preserve the functional relationships encoded by the one-shot demonstration. The process consists of *VLM-Based Scene Understanding*, *Object-Centric Geometric Anchoring*, and *Object-State-Aware Adaptation*.

1) VLM-Based Scene Understanding: Given the task description $\mathcal{T}$, we extract task-relevant prompts $p_k{}_{k=1}^{K}$ and query pretrained vision-language and vision foundation models, such as Florence-2 [17] and SAM2 [75], to obtain semantic object masks $\mathbf{M}_k^{\texttt{2D}}$ from the current observation $\mathcal{O}^{\texttt{new}}$ of each target object o_k. The masks establish explicit correspondence between language-described targets and their visual instances, allowing the system to disambiguate relevant objects in cluttered scenes. We further reconstruct a 3D dense point cloud $\mathcal{P}^{\texttt{3D}}$ using calibrated stereo matching algorithms [86], [87] and lift each mask to an object-level point cloud $\mathcal{P}_k^{\texttt{3D}}$. This provides the semantic and geometric information required for subsequent adaptation without task-specific detectors, CAD models, or additional training. The strong out-of-distribution perception capability of modern foundation models also makes this grounding procedure naturally robust to changes in unexpected illumination and distractors.

2) Object-Centric Geometric Anchoring: Instead of estimating a complete 6-DoF pose for every object [88], [89] or generating grasp proposals [90], [91], VLBiMan++ constructs compact task-relevant anchors according to object type. Examples of defined anchor points are shown in Fig. 3.

For ***rigid objects***, we use either the 2D mask centroid or a task-specific point on the foremost object–table contact boundary as the representative point. Given the corresponding 3D positions $\mathbf{p}_k^{\texttt{demo}}$ and $\mathbf{p}_k^{\texttt{new}}$, the positional variation is $\Delta\mathbf{x}_k = \mathbf{p}_k^{\texttt{new}} - \mathbf{p}_k^{\texttt{demo}}$. For orientation-sensitive objects,

Algorithm 1 Orientation Estimation for Rigid Objects

Require: Binary object mask $\mathbf{M}^{\texttt{2D}} \in \{0,1\}^{H\times W}$
Ensure: Orientation angle $\theta \in [0, 360)$ (degrees)
Extract object contour $C = \{(x_i, y_i)\}_{i=1}^{N}$ from $\mathbf{M}^{\texttt{2D}}$
Compute the centroid point $(\bar{x}, \bar{y})$ using image moments: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
Calculate second-order central moments: $\mu_{20} = \frac{1}{N}\sum(x_i - \bar{x})^2$, $\mu_{02} = \frac{1}{N}\sum(y_i - \bar{y})^2$, $\mu_{11} = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y})$
Construct covariance matrix: $\Sigma = \begin{bmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{bmatrix}$
Compute eigenvalues $\lambda_1 > \lambda_2$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of Σ
Obtain principal axis direction $\mathbf{a} = (a_x, a_y) = \mathbf{v}_1$
Project contour points onto principal axis:

$$p_i = (x_i - \bar{x})a_x + (y_i - \bar{y})a_y, \quad \forall i \in [1, N]$$

Identify endpoints: $e_{\text{max}} = \arg\max_i p_i$, $e_{\text{min}} = \arg\min_i p_i$
Calculate perpendicular width w_j within radius r around each e_j
Determine front endpoint: $e_{\text{front}} \leftarrow (w_{\text{max}} < w_{\text{min}})?e_{\text{max}} : e_{\text{min}}$
Adjust axis direction: $\mathbf{a} \leftarrow (\mathbf{a} \cdot (e_{\text{front}} - (\bar{x}, \bar{y})) < 0)? - \mathbf{a} : \mathbf{a}$
Final orientation angle: $\theta = \left(\arctan 2(a_y, a_x) \times \frac{180}{\pi}\right) \bmod 360$
return θ

we further extract the principal axis from the 2D mask using the image-moment method in Alg. 1, and compute $\Delta\theta_k = \angle(\mathbf{v}_k^{\texttt{new}}, \mathbf{v}_k^{\texttt{demo}})$. Specifically, the first-order moments determine the mask centroid, while the second-order image moments [92], [93] define a covariance matrix whose dominant eigenvector provides the principal axis. The resulting $(\Delta\mathbf{x}_k, \Delta\theta_k)$ is used to transform the demonstrated grasp or alignment pose into the current robot frame through the calibrated hand–eye transformation. Importantly, orientation estimation is invoked only when the manipulation is direction-sensitive. And symmetric objects with task-prescribed configurations require only positional anchoring.

For ***articulated objects***, the complete object may undergo configuration changes while the manipulation is still governed by a specific functional part. We therefore establish an explicit *body-part association* [94], [95] between the semantic object and its task-relevant component. Let o_k denote the complete articulated object and r_k denote its relevant part. Once r_k is identified, the latter is treated as the effective manipulation target, while the corresponding local geometry is regarded as a rigid entity for geometric anchoring. The same rigid-object procedure above, including position and orientation estimation when required, can therefore be applied to the local functional part without recovering the full articulated configuration. An example of an articulated penbag is shown in Fig. 3.

For ***deformable objects***, complete pose estimation is insufficient to represent the task-relevant geometry. Given the inherent complexity of modeling high-dimensional deformations, we focus on two prevalent topological categories: approximately *rectangular* objects (*e.g.*, towels) and *linear* objects (*e.g.*, cables). Instead of estimating a complete deformation state, we characterize the visible object by leveraging category-specific geometric priors to extract stable manipulation anchors. After morphologically smoothing the binary mask $\mathbf{M}^{\texttt{2D}}$ of a segmented deformable object to eliminate sensor noise [96], the algorithm routes the extraction process based on its topology. For *rectangular* objects, we enforce shape convexity

Algorithm 2 Anchors Extraction for Deformable Objects

Require: Binary object mask $\mathbf{M}^{\texttt{2D}} \in \{0,1\}^{H\times W}$, Object prior $c \in \{\texttt{Rectangular}, \texttt{Linear}\}$, Anchors number K
Ensure: Ordered anchors sequence $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$
Phase 1: Morphological Preprocessing
$\mathbf{M}^{\texttt{2D}}_{\text{closed}} \leftarrow \text{MorphClose}(\mathbf{M}^{\texttt{2D}}, \text{kernel})$
$\mathbf{M}^{\texttt{2D}}_{\text{smooth}} \leftarrow \text{Threshold}(\text{GaussianBlur}(\mathbf{M}^{\texttt{2D}}_{\text{closed}}, \sigma))$
Phase 2 & 3: Category-Specific Extraction
if $c == \texttt{Rectangular}$ **then**
  Extract largest external contour $\mathcal{C}$ from $\mathbf{M}^{\texttt{2D}}_{\text{smooth}}$
  Compute convex hull $\mathcal{H} \leftarrow \text{ConvexHull}(\mathcal{C})$
  $\mathcal{V} \leftarrow \text{ApproxPolyDP}(\mathcal{H}, \epsilon)$ ▷ Approximate to 4 vertices
  Compute centroid $\bar{\mathbf{v}} = \frac{1}{4}\sum_{\mathbf{v}\in\mathcal{V}} \mathbf{v}$
  Sort $\mathbf{v}_i \in \mathcal{V}$ by polar angle $\arctan 2(\mathbf{v}_{i,y} - \bar{\mathbf{v}}_y, \mathbf{v}_{i,x} - \bar{\mathbf{v}}_x)$
  Align starting node: shift $\mathcal{V}$ s.t. $\mathbf{v}_1 = \arg\min_{\mathbf{v}} \|\mathbf{v}\|_2$
  $\mathcal{P} \leftarrow \mathcal{V}$ ▷ Output 4 ordered corners
else if $c == \texttt{Linear}$ **then**
  $\mathbf{S} \leftarrow \text{Skeletonize}(\mathbf{M}^{\texttt{2D}}_{\text{smooth}})$
  Build undirected graph $G = (V, E)$ from skeleton pixels $\mathbf{S}$
  $\mathcal{J} \leftarrow \{v \in V \mid \deg(v) \geq 3\}$ ▷ Identify junction nodes
  for each junction cluster $J \subseteq \mathcal{J}$ **do**
    Compute cluster centroid $\mathbf{c}_J$, identify exit branches E_J
    Obtain lookahead vector $\vec{d}_e$ for each $e \in E_J$
    Greedy pair (e_i, e_j) if internal angle $\langle \vec{d}_{e_i}, \vec{d}_{e_j} \rangle \to \pi$
    Update $G \leftarrow (G \setminus J) \cup \{(e_i, e_j)\}$ ▷ Decouple crossings
  end for
  Find principal path $\Pi = \{\pi_1, \ldots, \pi_M\}$ on G
  Densify & smooth $\tilde{\Pi} \leftarrow \text{GaussianFilter}(\text{Interpolate}(\Pi), \sigma_{\text{path}})$
  Compute cumulative arc length $L(m) = \sum_{j=1}^{m} \|\tilde{\pi}_j - \tilde{\pi}_{j-1}\|_2$
  for $k \leftarrow 1$ to K **do**
    Target length $l_k = \frac{k-1}{K} \cdot L(|\tilde{\Pi}|)$
    $\mathbf{p}_k \leftarrow \tilde{\pi}_{\bar{m}}$, $\bar{m} = \arg\min_m |L(m) - l_k|$ ▷ Quantile
  end for
end if
return $\mathcal{P}$

on the external contour and employ a polygonal approximation to isolate four primary vertices. A centroid-based polar sorting mechanism is then applied to guarantee a consistent, ambiguity-free cyclic ordering of corners. For *linear* objects, we extract the medial-axis skeleton and construct an undirected topological graph. To robustly handle self-intersections, we cluster high-degree junction nodes and geometrically decouple them by greedily pairing exit branches based on their macroscopic tangent vectors. The continuous principal curve is subsequently recovered via Eulerian path tracing or Dijkstra's diameter search [97], enabling precise fractional sampling (*i.e.*, quantile locations) along the smoothed trajectory. This topology-aware extraction procedure retains critical task-relevant shape information with high efficiency, which is summarized in Alg. 2. And a broader visualization of anchors extraction results are presented in Fig. 4.

3) Object-State-Aware Adaptation: Let s_k denote the task-relevant object state o_k. Given the demonstrated state $s_k^{\texttt{demo}}$ and the current state $s_k^{\texttt{new}}$, we define the state variation as $\Delta s_k = s_k^{\texttt{new}} - s_k^{\texttt{demo}}$, where the subtraction denotes the representation-specific difference, such as translational displacement, angular deviation, articulation change, or boundary-anchor variation. VLBiMan++ selectively updates only state-conditioned skills:

$$\widetilde{\mathcal{M}}_i^{\texttt{SI}} = \mathcal{M}_i^{\texttt{SI}}, \quad \widetilde{\mathcal{M}}_j^{\texttt{SC}} = \mathcal{A}(\mathcal{M}_j^{\texttt{SC}}, \Delta s_k), \tag{6}$$

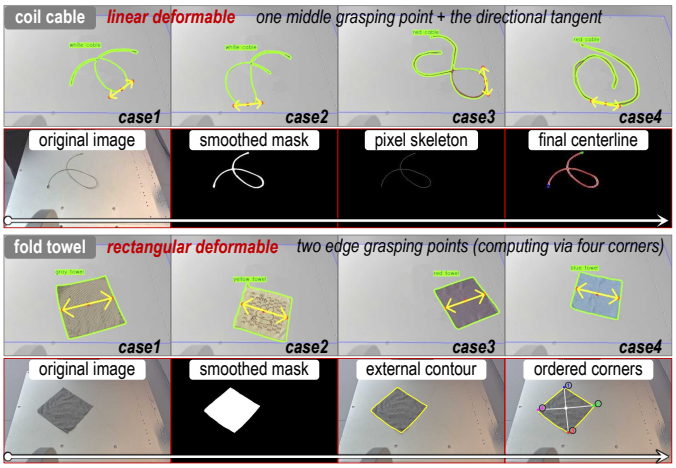


Fig. 4. Illustrations of extracted various anchor points for deformable objects in two tasks including `coil cable` (top) and `fold towel` (bottom). These can serve as a supplement to Alg. 2. We present the computational results for four test cases of each task. In addition, we also present intermediate results for handling linear or rectangular deformable objects.

where $\mathcal{A}(\cdot)$ denotes the corresponding anchor-based adaptation operation. For rigid objects, Δs_k is primarily reflected by $(\Delta \mathbf{x}_k, \Delta\theta_k)$ and size variation. For articulated objects, it additionally contains the state of the relevant movable part. For deformable objects, it is represented by the variation of the extracted boundary anchors. Thus, VLBiMan++ preserves invariant manipulation structure while adapting only the components affected by object-state changes, and passes the resulting skills to the trajectory composition stage.

D. Autonomous Trajectory Composition

After state-aware adaptation, VLBiMan++ composes the state-invariant and state-conditioned skills according to the temporal structure of the demonstration. Since direct concatenation may introduce reachability or collision issues under novel object states, we apply progressive IK refinement and dynamic collision compensation, followed by closed-loop re-observation and recomposition during execution.

1) Progressive IK Refinement: For the initial grasping stage, directly solving IK from the current end-effector pose to the adapted grasp pose may yield unreachable or unsafe motions. VLBiMan++ therefore approaches the target through interpolated Cartesian waypoints:

$$\begin{aligned} \mathbf{q}^{(n+1)} &= \mathtt{IK}(\mathbf{T}_g^{(n)}), \\ \mathbf{T}_g^{(n)} &= \mathtt{SplineInterp}(\mathbf{T}_{\mathtt{start}}, \mathbf{T}_{\mathtt{goal}}, n), \end{aligned} \tag{7}$$

where $\mathbf{T}_{\mathtt{start}}$ is the currently observed end-effector pose, $\mathbf{T}_{\mathtt{goal}}$ is the adapted grasp pose obtained from Sec. III-C, and n is the interpolation density. This converts a large reaching motion into a sequence of smaller and more feasible IK subproblems [84], [85], improving reachability and providing intermediate states for object-state re-evaluation. If the object is displaced during this process, $\mathbf{T}_{\mathtt{goal}}$ is recomputed from the current observation. Otherwise, it remains unchanged.

2) Dynamic Collision Compensation: To reduce premature contact during pre-grasping, we modify the adapted target position with proximal and vertical compensation:

$$\widetilde{\mathbf{x}}^{\mathtt{goal}} = \mathbf{x}^{\mathtt{goal}} + \delta_{\mathtt{base}}\mathbf{u}_{\parallel} + \delta_z \mathbf{u}_z, \tag{8}$$

where $\delta_{\mathtt{base}}$ and δ_z denote the proximal and vertical compensation terms, respectively, and $\mathbf{u}_{\parallel}$ and $\mathbf{u}_z$ specify the corresponding Cartesian directions. The proximal compensation increases clearance along the horizontal approach direction, while the vertical compensation provides additional separation for top-down reaching. After trajectory synthesis, a one-time physical replay is used to identify unintended object or inter-arm collisions, after which the corresponding waypoints are adjusted before reuse. Thus, collision-related corrections are incorporated into the reusable skill rather than repeatedly optimized during subsequent deployments.

3) Closed-Loop Re-observation and Re-composition: Real-world execution may deviate from the planned state because of external perturbations or robot–object interactions. VLBiMan++ therefore monitors the task-relevant object state and triggers re-adaptation when $|s_k^{\mathtt{exec}} - s_k^{\mathtt{plan}}| > \epsilon$, where $s_k^{\mathtt{exec}}$ and $s_k^{\mathtt{plan}}$ denote the observed and planned states, respectively. The ϵ denotes the stability tolerance used to distinguish effective object displacement from perceptual fluctuations. Once triggered, the target is re-grounded and the affected state-conditioned skills are re-instantiated:

$$\widetilde{\mathcal{M}}_j^{\mathtt{SC}} \leftarrow \mathcal{A}\left(\mathcal{M}_j^{\mathtt{SC}}, s_k^{\mathtt{exec}} - s_k^{\mathtt{demo}}\right), \quad \widetilde{\mathcal{M}}_i^{\mathtt{SI}} \leftarrow \mathcal{M}_i^{\mathtt{SI}}. \tag{9}$$

The resulting trajectory is then recomposed while preserving all unaffected invariant skills. This selective closed-loop update enables VLBiMan++ to withstand repeated object relocation and execution-induced displacement without retraining, following the cycle *Observe* → *Ground* → *Adapt* → *Compose* → *Execute* → *Re-observe*. This design preserves the reusable task prior extracted from a single demonstration while enabling sustained operation under changing object states and external disturbances, forming the basis for deployment-level generalization investigated in this work.

IV. Experiments

We aim to answer four key questions: (1) How effectively does VLBiMan++ perform diverse bimanual manipulation tasks without retraining (Sec. IV-B)? (2) How far can a single demonstration generalize across tasks, objects, scenes, and robot embodiments (Sec. IV-C)? (3) How robustly can VLBiMan++ sustain manipulation under dynamic disturbances and long-term closed-loop execution (Sec. IV-D)? (4) How do individual components contribute to the performance, generalization, and robustness of the system (Sec. IV-E)? We evaluate VLBiMan++ on diverse real-world bimanual manipulation tasks using heterogeneous robotic platforms.

A. Experimental Setup

1) Tasks and Robotic Platforms: We evaluate VLBiMan++ on a substantially expanded suite of real-world bimanual manipulation tasks covering diverse skill compositions, object states, and environmental conditions (refer Tab. I). The core benchmark retains the ten tasks from our conference version (refer Fig. 5 and Fig. 7), including six primary bimanual

TABLE I

DETAILED STATISTICS ON ALL BIMANUAL MANIPULATION TASKS, INCLUDING 10 PREVIOUS TASKS AND 9 NEWLY ADDED TASKS. IN THE **DUAL-ARM PLATFORM(S)**, P1 AND P2 REPRESENT THE TASK HAS BEEN INSTANTIATED ON A CONTRALATERAL ROBOT AND A HUMANOID ROBOT, RESPECTIVELY. IN THE **OBJECT PROPERTIES**, RIG., ART., AND DEF. REPRESENT RIGID, ARTICULATED, AND DEFORMABLE OBJECTS, RESPECTIVELY. AMONG THEM, DEFORMABLE OBJECTS REQUIRE ALG. 2 TO EXTRACT ANCHOR POINTS. AND THE **ORIENTATION ESTIMATION** IS BASED ON ALG. 1.

	primary bimanual tasks/skills						*long-horizon multi-stage tasks*				*new rearrangement tasks*						*new non-rigid targets*		
Task Types & Names	`plugpen`	`inserting`	`unscrew`	`pouring`	`pressing`	`handover`	`reorient +unscrew`	`unscrew +pouring`	`tool-use scoop`	`tool-use funnel`	`flatting`	`reorient`	`flipping`	`upright`	`place bottle_mug`	`place fork_spoon`	`zip penbag`	`coil cable`	`fold towel`
Dual-Arm Platform(s)	P1|P2	P1|P2	P1|P2	P1|P2	P1	P1|P2	P1	P1|P2	P1	P1	P2	P2	P2	P2	P2	P2	P2	P2	P2
Object Number	2	2	1	2	2	1	1	2	3	3	1	1	1	1	2	2	2	1	1
Object Names /States	pen body +cap	pen +cup	upward bottle	upward bottle +mug	cup+ pump bottle	spoon or shovel	lying bottle	upward bottle +mug	spoon+ bowl_1+ bowl_2	funnel +bottle +mug	upside down bottle	lying bottle	upside down mug	lying mug	bottle +mug	fork+ spoon	penbag +pen	linear cable	rectan -gular towel
Object Properties	Art.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Rig.	Art.	Def.	Def.
Orientation Estimation?	✓✓	✓✗	✗	✗✗	✗✗	✓	✓	✗✗	✓✗✗	✓✗✗	✗	✓	✗	✓	✓✓	✓✓	✓✓	✗	✗

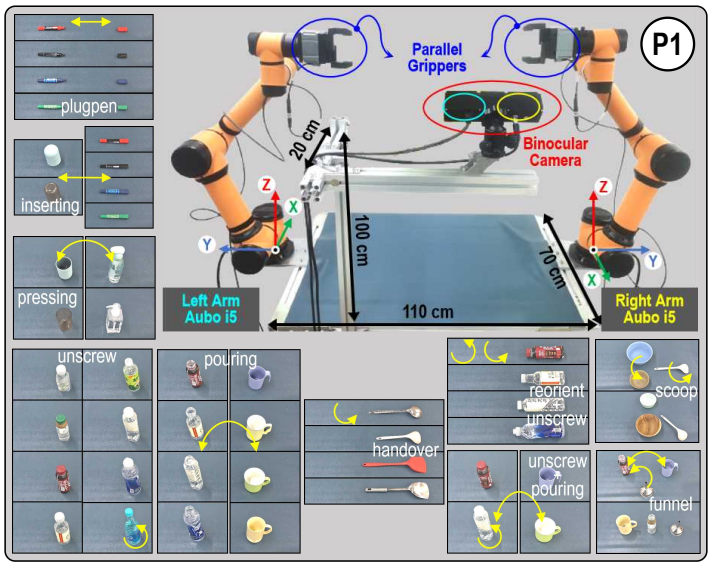


Fig. 5. The contralateral dual-arm platform (**P1**), and all manipulated object assets involved in ten previously defined bimanual tasks.

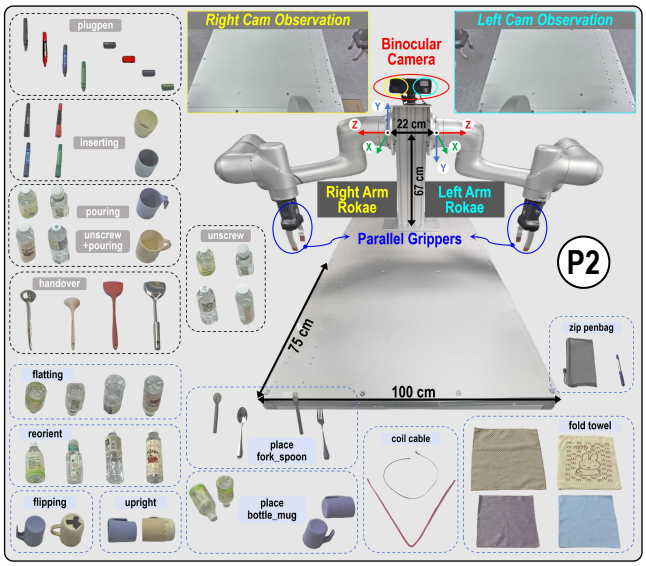


Fig. 6. The humanoid platform (**P2**), and all manipulated object assets involved in six previous and nine newly defined bimanual tasks.

tasks/skills[1] and four more challenging long-horizon or tool-use tasks. In this journal extension, we further introduce nine newly-defined tasks and object configurations to probe generalization beyond original rigid-object settings, including unseen object categories, articulated and deformable objects, cluttered scenes, and prolonged disturbance-aware execution (refer Fig. 6 and Fig. 8). The experiments are conducted on two heterogeneous dual-arm platforms: a contralateral fixed-base setup equipped with two 6-DoF Aubo-i5 arms[2] (880mm reach, denoted as **P1**), parallel grippers, and a third-person stereo camera, and a humanoid-style platform equipped with two 6-DoF Rokae xMate CR7 arms[3] (988mm reach, denoted as **P2**), parallel grippers, and a centrally mounted stereo camera. This setup enables evaluation across diverse tasks, object states, scenes, and robot embodiments.

2) Baselines and Evaluation Metrics: We compare VLBiMan++ with representative training-free and data-efficient manipulation approaches, including Robot-ABC [36], which combines affordance prediction with AnyGrasp [91] for grasp generation, and ReKep [24], which leverages vision foundation models (SAM [98] and DINOv2 [74]) and LLM-guided keypoint constraints for task execution (*e.g.*, GPT-4o [15]). We further include one-shot manipulation methods such as Mechanisms [61] and MAGIC [62], adapted to our bimanual setting where applicable, as well as the enhanced ReKep+ variant with oracle initial grasp annotations for a stronger comparison. For all real-robot evaluations, we report task success rate over repeated trials under matched experimental conditions, considering both seen and unseen object instances as well as external disturbances when applicable. Additional analyses evaluate execution efficiency, robustness to clutter and repeated perturbations, and cross-embodiment transfer, providing a comprehensive assessment of VLBiMan++ across

[1]We have renamed the inherently dual-arm task `reorient` from the original conference version to `handover` to avoid confusion with the newly added single-arm (left or right) task `reorient`.

[2]https://www.aubo-cobot.com/public/i5product3

[3]https://www.rokae.com/en/product/show/545/xMateCR.html

TABLE II
QUANTITATIVE COMPARISON RESULTS OF SUCCESS RATES ON SIX PRIMARY BIMANUAL TASKS/SKILLS AND FOUR LONG-HORIZON MULTI-STAGE TASKS UNDER THE **IN-DISTRIBUTION (ID)** SETTING (NEW PLACEMENTS + **SAME OBJECTS**) OR THE **OUT-OF-DISTRIBUTION (OOD)** SETTING (NEW PLACEMENTS + **NOVEL INSTANCES**) USING THE CONTRALATERAL DUAL-ARM ROBOT (**P1**).

Test Setting	Dyna-mic Inter-ference	Manipulation Method	*six primary bimanual tasks/skills*							*four long-horizon multi-stage tasks*				
			plugpen	inserting	unscrew	pouring	pressing	handover	Average Success Rate	reorient +unscrew	unscrew +pouring	tool-use scoop	tool-use funnel	Average Success Rate
ID	No	Mechanisms	11/25	09/25	05/25	05/25	07/25	03/25	26.7%	05/25	04/25	02/25	01/25	12.0%
		MAGIC	16/25	15/25	10/25	10/25	09/25	07/25	44.7%	09/25	08/25	04/25	03/25	24.0%
		Robot-ABC	14/25	10/25	09/25	07/25	08/25	06/25	36.0%	06/25	06/25	03/25	03/25	18.0%
		ReKep	14/25	11/25	10/25	12/25	10/25	08/25	43.3%	07/25	08/25	05/25	03/25	23.0%
		ReKep+	19/25	18/25	13/25	17/25	17/25	11/25	63.3%	11/25	10/25	07/25	06/25	34.0%
		VLBiMan++	25/25	23/25	20/25	21/25	20/25	19/25	**85.3%**	15/25	15/25	12/25	10/25	**52.0%**
	Yes	Mechanisms	05/25	05/25	03/25	02/25	04/25	01/25	13.3%	01/25	02/25	00/25	00/25	3.0%
		MAGIC	09/25	09/25	05/25	04/25	06/25	04/25	24.7%	04/25	03/25	04/25	01/25	12.0%
		Robot-ABC	07/25	06/25	04/25	03/25	05/25	02/25	18.0%	02/25	02/25	02/25	02/25	9.0%
		ReKep	10/25	06/25	06/25	04/25	05/25	03/25	22.7%	06/25	05/25	03/25	02/25	16.0%
		ReKep+	12/25	10/25	09/25	08/25	09/25	09/25	38.0%	08/25	08/25	05/25	03/25	24.0%
		VLBiMan++	19/25	16/25	19/25	18/25	17/25	15/25	**69.3%**	12/25	11/25	09/25	06/25	**38.0%**
OOD	No	Mechanisms	06/25	05/25	02/25	01/25	04/25	01/25	12.7%	01/25	02/25	00/25	00/25	3.0%
		MAGIC	11/25	10/25	05/25	05/25	06/25	04/25	27.3%	05/25	04/25	01/25	01/25	11.0%
		Robot-ABC	11/25	09/25	03/25	02/25	07/25	04/25	24.0%	04/25	02/25	00/25	01/25	7.0%
		ReKep	12/25	08/25	05/25	06/25	07/25	06/25	29.3%	05/25	04/25	01/25	00/25	10.0%
		ReKep+	15/25	12/25	09/25	10/25	11/25	07/25	42.7%	07/25	06/25	04/25	02/25	19.0%
		VLBiMan++	24/25	21/25	18/25	17/25	20/25	17/25	**78.0%**	12/25	11/25	10/25	08/25	**41.0%**
	Yes	Mechanisms	03/25	01/25	00/25	00/25	02/25	00/25	4.0%	00/25	00/25	00/25	00/25	0.0%
		MAGIC	05/25	04/25	03/25	01/25	04/25	01/25	12.0%	02/25	02/25	01/25	00/25	5.0%
		Robot-ABC	05/25	03/25	00/25	00/25	03/25	00/25	7.3%	01/25	00/25	01/25	00/25	2.0%
		ReKep	09/25	04/25	03/25	01/25	04/25	02/25	15.3%	03/25	03/25	00/25	00/25	6.0%
		ReKep+	10/25	08/25	05/25	04/25	06/25	05/25	25.3%	06/25	04/25	01/25	01/25	12.0%
		VLBiMan++	18/25	14/25	15/25	14/25	15/25	13/25	**59.3%**	08/25	09/25	05/25	02/25	**24.0%**

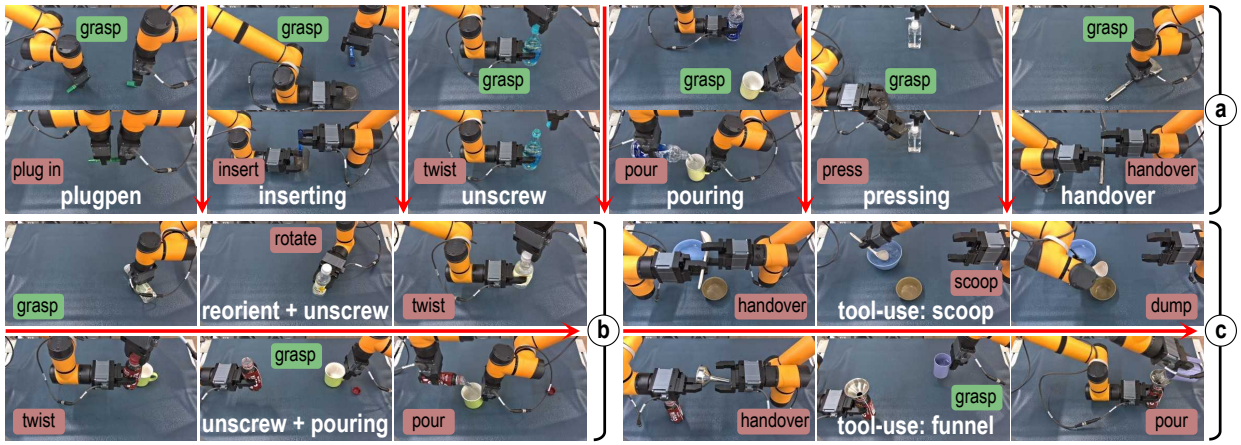


Fig. 7. Visualizations of ten tasks executed on the contralateral dual-arm robots platform **P1**. They are designed to validate different aspects, including **(a)** six dual-arm primary skills, **(b)** combination of basic skills for two long-horizon tasks, and **(c)** exploration of two multi-stage tool-use tasks.

the proposed generalization boundary.

B. Overall Performance of VLBiMan++

We first evaluate the overall capability of VLBiMan++ via a substantially expanded real-world tests. In addition to the ten tasks inherited from VLBiMan (mainly on the contralateral platform **P1**), the journal version further introduces six object-rearrangement tasks and three non-rigid manipulation tasks (mainly on the humanoid platform **P2**), covering a broader range of task structures and object states. For each task, we evaluate both in-distribution (ID) and out-of-distribution (OOD) settings, including novel object instances and spatial configurations, and report the corresponding success rates in Tab. II and Tab. III. Unless otherwise specified, each setting is evaluated over 25 real-robot trials.

1) General Bimanual Manipulation: As summarized in Tab. II and Tab. III, VLBiMan++ maintains the strong performance of the original VLBiMan while substantially broadening the scope of evaluated manipulation behaviors. The system consistently achieves reliable execution on foundational bimanual tasks as well as long-horizon and tool-use tasks, and further extends to the newly introduced rearrangement and non-rigid manipulation settings. This indicates that the task-aware decomposition and object-state-aware adaptation introduced in VLBiMan++ are not restricted to a small set of rigid-object behaviors, but can support more diverse task structures and object conditions. Notably, the same one-shot demonstration is reused as the task prior across the corresponding task variations, while only state-dependent components are adapted to the current scene. The resulting trajectories preserve

TABLE III

QUANTITATIVE COMPARISON RESULTS OF SUCCESS RATES ON SIX NEW REARRANGEMENT TASKS AND THREE NEW NON-RIGID TARGETS UNDER THE **IN-DISTRIBUTION (ID)** SETTING (NEW PLACEMENTS + **SAME OBJECTS**) OR THE **OUT-OF-DISTRIBUTION (OOD)** SETTING (NEW PLACEMENTS + **NOVEL INSTANCES**) USING THE HUMANOID DUAL-ARM ROBOT (**P2**). THE '—' MEANS THERE ARE NO EXTRA NEW OBJECT INSTANCES.

Test Setting	Dynamic Interference	Manipulation Method	*six new rearrangement tasks*							*three new non-rigid targets*			
			flatting	reorient	flipping	upright	place bottle_mug	place fork_spoon	Average Success Rate	zip penbag	coil cable	fold towel	Average Success Rate
ID	No	Mechanisms	04/25	04/25	02/25	01/25	01/25	01/25	8.7%	00/25	00/25	02/25	2.7%
		MAGIC	06/25	05/25	03/25	02/25	01/25	02/25	12.7%	00/25	00/25	03/25	4.0%
		Robot-ABC	07/25	06/25	06/25	06/25	05/25	05/25	23.3%	01/25	00/25	05/25	8.0%
		ReKep+	13/25	12/25	10/25	11/25	08/25	11/25	43.3%	04/25	00/25	08/25	16.0%
		VLBiMan++	24/25	22/25	18/25	20/25	15/25	23/25	**81.3%**	12/25	10/25	20/25	**56.0%**
	Yes	Mechanisms	02/25	02/25	01/25	00/25	00/25	01/25	4.0%	00/25	00/25	00/25	0.0%
		MAGIC	02/25	03/25	01/25	01/25	00/25	02/25	6.0%	00/25	00/25	00/25	0.0%
		Robot-ABC	05/25	05/25	03/25	03/25	01/25	04/25	14.0%	01/25	00/25	02/25	4.0%
		ReKep+	08/25	08/25	06/25	05/25	03/25	08/25	25.3%	02/25	00/25	05/25	9.3%
		VLBiMan++	15/25	14/25	13/25	13/25	11/25	15/25	**54.0%**	06/25	02/25	13/25	**28.0%**
OOD	No	Mechanisms	04/25	04/25	01/25	01/25	—	01/25	8.8%	—	00/25	01/25	2.0%
		MAGIC	07/25	06/25	03/25	04/25	—	02/25	17.6%	—	00/25	03/25	6.0%
		Robot-ABC	09/25	08/25	04/25	06/25	—	03/25	24.0%	—	00/25	04/25	8.0%
		ReKep+	12/25	10/25	06/25	09/25	—	05/25	33.6%	—	00/25	06/25	12.0%
		VLBiMan++	23/25	20/25	15/25	18/25	—	20/25	**60.8%**	—	06/25	14/25	**40.0%**
	Yes	Mechanisms	02/25	01/25	00/25	00/25	—	00/25	2.4%	—	00/25	00/25	0.0%
		MAGIC	04/25	03/25	02/25	02/25	—	01/25	9.6%	—	00/25	00/25	0.0%
		Robot-ABC	05/25	04/25	02/25	02/25	—	01/25	11.2%	—	00/25	00/25	0.0%
		ReKep+	07/25	06/25	03/25	04/25	—	02/25	19.2%	—	00/25	01/25	2.0%
		VLBiMan++	16/25	12/25	09/25	11/25	—	15/25	**50.4%**	—	01/25	08/25	**18.0%**

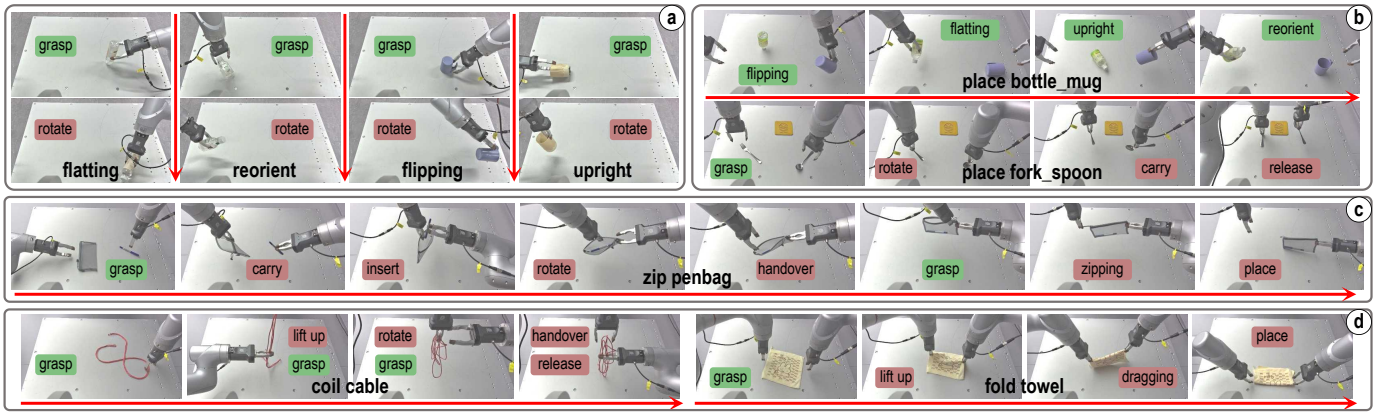


Fig. 8. Visualizations of newly added nine tasks executed on the humanoid-style dual-arm platform **P2**. They are defined to further confirm the universality of VLBiMan++, including **(a)** four left-arm or right-arm rearrangement skills, **(b)** combination of new basic skills for two long-horizon rearrangement tasks, **(c)** a complex task involving articulated bodies, **(d)** two dual-arm tasks involving linear or rectangular deformable objects.

the essential functional relationships of the demonstrated behaviors, including precise grasping, object alignment, inter-arm coordination, and multi-stage skill composition. The consistently high performance across both ID and OOD settings therefore demonstrates that VLBiMan++ improves not only task coverage, but also the practical reusability of a single human demonstration.

2) Comparison with Existing Methods: We further compare VLBiMan++ with representative state-of-the-art methods under the same real-world evaluation protocol. As shown in Tabs. II and III, VLBiMan++ achieves the most competitive overall success rates while requiring only a single human demonstration and no policy retraining. In contrast, data-driven imitation and VLA policies [8], [13], [14], [18] typically require substantial task-specific robot demonstrations and additional optimization when adapting to new tasks or objects, whereas existing training-free modular methods such as ReKep [24] and Robot-ABC [36] must regenerate task-dependent grasps or trajectories for each new configuration. The adapted another two baselines Mechanisms [61] and MAGIC [62] originally designed for single-arm tasks also cannot handle these bimanual tasks well, revealing the non-trivial nature of dual-arm coordination. This comparison highlights an important distinction in evaluation philosophy: VLBiMan++ is designed for one-shot, training-free generalization, rather than maximizing performance through additional task-specific data. Within this setting, its ability to reuse invariant skills, selectively adapt state-dependent components, and operate across novel objects and scenes provides a favorable balance between success rate, demonstration efficiency, and deployment cost. The advantage becomes more pronounced as task complexity, object diversity, and environmental variation increase, where repeatedly reconstructing complete manipulation trajectories becomes increasingly inefficient. Overall, these results demonstrate that VLBiMan++ preserves the effectiveness of VLBiMan while substantially extending its applicability to a broader range

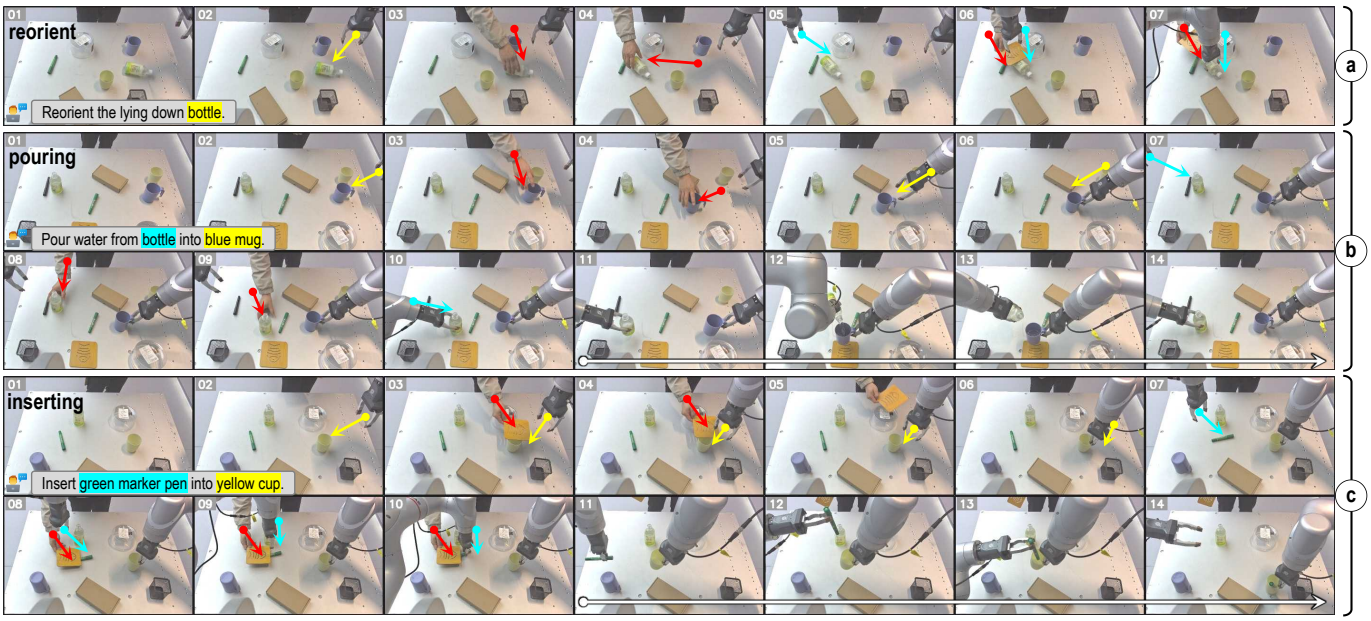


Fig. 9. Visualizations of VLBiMan++'s robustness performance in cluttered scenarios on the platform **P2**. We selected three tasks for extensive evaluation: (**a**) `reorient`, (**b**) `pouring`, and (**c**) `inserting`. The red arrow indicates the direction of the manually moved or deliberately obscured object (interfering). The cyan arrow and yellow arrow indicate the movement direction of the left and right robotic arms (chasing) respectively.

of bimanual manipulation problems, establishing a strong foundation for the subsequent analyses of task, object, scene, embodiment, and deployment generalization.

C. Generalization across Different Configurations

We next evaluate how far the one-shot manipulation prior can be reused under increasingly diverse task, object, scene, and embodiment setttings. The results establish four intermediate levels of the generalization boundary introduced in VLBiMan++: task and skill composition, object-state variation, cluttered scenes, and heterogeneous robotic embodiments.

1) Task and Skill Generalization: We first investigate whether a single demonstration can provide reusable building blocks beyond the original task sequence. Starting from six primary bimanual tasks, VLBiMan++ reuses extracted atomic skills to construct long-horizon behaviors, novel combinations, and tool-use tasks. In particular, `reorient+unscrew`, `unscrew+pouring`, `place bottle_mug` and `place fork_spoon` compose previously acquired skills, while `tool-use spoon` and `tool-use funnel` further require their integration across multiple objects and stages. As shown in Tabs. II and III, VLBiMan++ maintains reliable performance across these increasingly complex settings without collecting additional demonstrations, demonstrating the transition from one demonstration to reusable skills and ultimately to new task compositions.

2) Unseen Object and Object-State Generalization: We further examine the object-level generalization enabled by the *Object-State-Aware Adaptation* module in Sec. III-C. For rigid objects, we evaluate unseen categories as well as variations in shape, size, position, and orientation. VLBiMan++ transfers the demonstrated manipulation structure by adapting only the task-relevant geometric anchors. We then extend the evaluation to articulated objects (`zip penbag`), where changes in movable-part configurations require adaptation of the corresponding body-part relationships. Finally, we consider non-rigid objects, including deformable structures (`coil cable` and `fold towel`) like cables and towels, for which VLBiMan++ employs topology-aware boundary anchors rather than a single rigid pose. Across these settings, the demonstrated task prior remains reusable while only the state-dependent components are modified, showing that the proposed framework extends beyond conventional rigid-instance transfer toward broader object-state generalization.

3) Cluttered-Scene Generalization: To assess scene-level generalization, we further evaluate VLBiMan++ in cluttered tabletop environments containing multiple distractors, semantic ambiguity, partial occlusion, and potential spatial interference. We conduct additional experiments on `reorient`, `pouring`, and `inserting`, with at least five irrelevant objects introduced into each scene (the platform is **P2**). Some distractors are visually or semantically similar to the target objects, while others may interfere with the manipulation path. The same vision-language grounding and adaptation pipeline is retained without task-specific modification. As shown in Fig. 9, VLBiMan++ consistently identifies the language-specified targets, rejects irrelevant or similar objects, and maintains reliable manipulation despite partial occlusion and object relocation during execution. These results demonstrate that the object-centric anchors remain effective beyond clean tabletop configurations and that semantic grounding provides an important interface between cluttered visual observations and reusable manipulation skills.

4) Cross-Embodiment Transfer: Finally, we evaluate whether the task prior extracted from a single demonstration can be preserved when its physical realization changes across robotic embodiments. We transfer VLBiMan++ from the contralateral platform (**P1**) to a humanoid-style platform (**P2**) equipped with different robot arms,

TABLE IV
QUANTITATIVE RESULTS OF VLBIMAN++'S SUCCESS RATES ON **SIX TRANSFERRED BIMANUAL TASKS** FROM THE CONTRALATERAL DUAL-ARM ROBOT (**P1**) TO THE HUMANOID DUAL-ARM ROBOT (**P2**) UNDER THE **ID** AND **OOD** SETTINGS.

Dynamic Interference	Dual-Arm Platform Type	*ID setting: new placements + same objects*							*OOD setting: new placements + novel instances*						
		plugpen	inserting	unscrew	pouring	handover	unscrew +pouring	Average Success Rate	plugpen	inserting	unscrew	pouring	handover	unscrew +pouring	Average Success Rate
No	Contralateral	19/20	18/20	16/20	17/20	15/20	12/20	80.8%	15/20	17/20	14/20	14/20	14/20	09/20	69.2%
	Humanoid	20/20	19/20	17/20	16/20	15/20	13/20	83.3%	16/20	18/20	16/20	13/20	14/20	09/20	71.7%
Yes	Contralateral	16/20	13/20	15/20	15/20	12/20	08/20	65.8%	14/20	11/20	12/20	11/20	10/20	07/20	54.2%
	Humanoid	16/20	14/20	15/20	14/20	13/20	09/20	67.5%	12/20	12/20	13/20	12/20	10/20	08/20	55.8%

TABLE V
QUANTITATIVE COMPARISON RESULTS OF SUCCESS RATES ON **SIX PRIMARY BIMANUAL SKILLS/TASKS** USING THE CONTRALATERAL DUAL-ARM ROBOT (**P1**). ALL TASKS FACED THE CHALLENGES OF **DYNAMIC INTERFERENCE** AND **UNEVEN ILLUMINATION**.

Manipulation Method	*ID setting: new placements + same objects*							*OOD setting: new placements + novel instances*						
	plugpen	inserting	unscrew	pouring	pressing	handover	Average Success Rate	plugpen	inserting	unscrew	pouring	pressing	handover	Average Success Rate
Mechanisms	01/20	01/20	00/20	01/20	01/20	00/20	3.3%	00/20	00/20	00/20	00/20	00/20	00/20	0.0%
MAGIC	03/20	04/20	02/20	02/20	02/20	01/20	11.7%	01/20	02/20	00/20	01/20	01/20	00/20	4.2%
Robot-ABC	02/20	02/20	01/20	01/20	01/20	01/20	6.7%	00/20	00/20	00/20	00/20	00/20	00/20	0.0%
ReKep	05/20	03/20	02/20	01/20	02/20	02/20	12.5%	03/20	01/20	01/20	01/20	01/20	00/20	5.8%
ReKep+	08/20	06/20	04/20	03/20	04/20	05/20	25.0%	05/20	02/20	03/20	02/20	03/20	01/20	13.3%
VLBiMan++	14/20	13/20	14/20	15/20	14/20	11/20	**67.5%**	13/20	11/20	11/20	10/20	12/20	10/20	**55.8%**

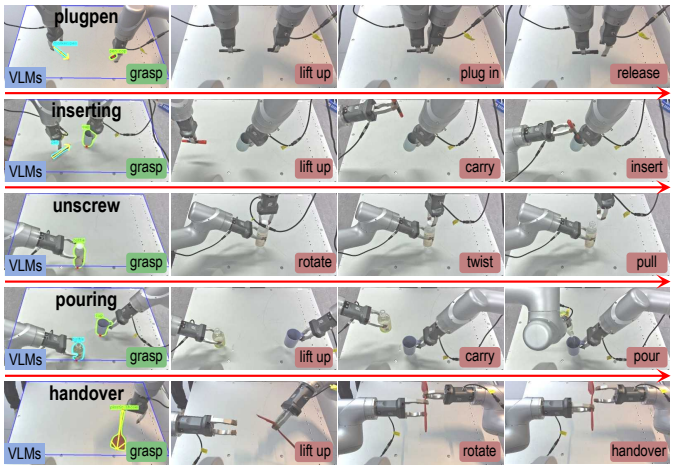


Fig. 10. Visualizations of five cross-embodiment transferred primary bimaual tasks executed on the new humanoid-style robot (**P1** → **P2**).

gripper specifications, and arm configurations. We evaluate `plugpen`, `inserting`, `unscrew`, `pouring`, `handover` and `unscrew+pouring`, including task configurations that are modified to accommodate the new embodiment. As summarized in Tab. IV, VLBiMan++ maintains competitive performance on both seen and unseen objects, with additional robustness under external perturbations. These results indicate that the demonstration prior is not tied to the original robot's absolute configuration; instead, its task-relevant structure can be re-instantiated on a different embodiment. Qualitative executions in Fig. 10 further show that the humanoid platform can exploit more synchronized dual-arm motions, yielding smoother and more human-like execution.

D. Robustness and Long-Term Closed-Loop Deployment

Beyond static generalization, we further evaluate whether VLBiMan++ can sustain reliable manipulation under continuously changing execution conditions. We focus on three aspects: robustness to dynamic disturbances, stability during prolonged operation, and the effectiveness of the closed-loop re-observation and re-composition mechanism.

1) Dynamic Disturbance Robustness: We first evaluate robustness to external disturbances by varying the number and type of perturbations applied during execution. In addition to object relocation, we consider partial occlusion and uneven illumination, which jointly challenge visual grounding and geometric adaptation. Specifically, we conduct experiments on all six primary bimanual tasks using the platform **P1**, focusing on the ID evaluations without loss of generality. We define one interference as a scenario where each object involved in the task is disturbed once. Under this definition, we systematically vary the number of interferences from 0 to 5 and record the average task success rates, which are: 85.0%, 70.0%, 61.7%, 56.7%, 53.3%, and 50.8%, respectively. As the disturbance frequency increases, the success rate decreases monotonically, but the degradation becomes progressively smaller, indicating a diminishing marginal impact of repeated perturbations. This behavior suggests that the progressively updated object-relative anchors remain effective as the end-effector approaches the target. VLBiMan++ also exhibits strong robustness to adverse visual conditions. Under uneven illumination, the binary-mask-based geometric representation remains largely unaffected, while the underlying VLM/VFM perception maintains reliable object grounding. As summarized in Tab. V, the resulting performance degradation is substantially smaller than that of the compared various baselines, demonstrating that the proposed perception–adaptation pipeline is robust not only to geometric disturbances but also to appearance variations. Some qualitative examples of anti-interference on the platform **P2** can be found in Fig. 9.

2) Super Long-Duration Execution: To further assess deployment stability, we conduct super long-duration tests using the platform **P2** in which the task `reorient` is executed

Fig. 11. Video snapshots recorded from a third-person perspective for the super long-duration experiment of `reorient` on platform **P2**. It contains keyframes from **20** trials in which the left arm (cyan circle) or right arm (yellow circle) was used to upright a bottle lying down in an arbitrary orientation.

continuously over **20** consecutive trials, with object states repeatedly randomized and external disturbances introduced throughout the process. Unlike conventional evaluations based on a limited number of independent trials, this setting examines whether small errors accumulate over time and whether the system can remain operational without retraining or manual re-initialization. The results show that VLBiMan++ maintains a stable success rate over prolonged execution, despite repeated object relocation and execution-induced disturbances. This stability is enabled by the selective reuse of state-invariant skills and repeated adaptation of only the affected components, preventing minor deviations from accumulating into a complete trajectory failure. The long-duration evaluation therefore provides evidence that the one-shot task prior remains reusable not only across isolated episodes, but also over sustained real-world deployment. We present the third-person video snapshots of this experiment in Fig. 11.

3) Closed-Loop Re-observation and Re-composition: We finally isolate the contribution of the closed-loop mechanism introduced in Sec. III-D by comparing execution with and without re-observation and re-composition under repeated object perturbations. When the observed object state remains within the tolerance ϵ (*e.g.*, 10 mm), VLBiMan++ continues the current trajectory without unnecessary replanning; once the discrepancy exceeds ϵ, the system re-identifies the target, updates the corresponding state-conditioned skills, and recomposes the remaining trajectory. This mechanism leads to several practically important behaviors. During pre-grasping, the robot can dynamically refine its approach after an object is relocated. When the target is temporarily partially occluded, execution can continue using the temporally structured trajectory and resume adaptation once reliable visual grounding is recovered. Moreover, small object displacements during intermediate manipulation stages can often be tolerated without interrupting the task. These observations are consistent with the closed-loop execution traces shown in Fig. 12 and demonstrate that VLBiMan++ is capable of selectively correcting its behavior rather than restarting the entire manipulation process.

E. Ablation and In-Depth Analysis

We finally investigate why VLBiMan++ works and where its current limitations arise. We evaluate the robustness of

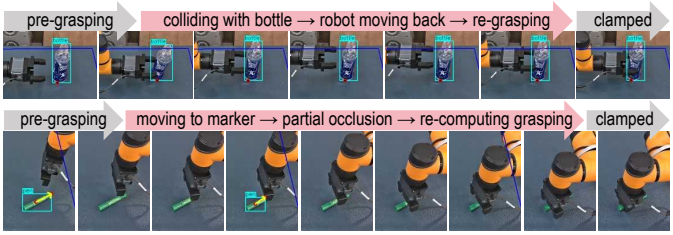


Fig. 12. Examples of interesting findings about the closed-loop mechanism. ***Top row***: this case comes from the pre-grasping phase of `pouring`, where the left arm approaches and grasps the bottle. ***Bottom row***: this case comes from the pre-grasping phase of `inserting`, where the right arm approaches and grasps the marker from the top direction.

object-centric representing points, the sensitivity to pre-grasp interpolation density, the contribution of individual modules, and the distribution of real-world failure cases.

1) Representative-Point Robustness: VLBiMan++ adopts two lightweight yet intuitive object-centric anchors: the 2D mask centroid and the foremost object-table contact point. Despite their simplicity, these anchors provide a stable interface for transferring task-relevant spatial relationships across unseen object geometries without object-specific models. We further evaluate their robustness in the `inserting` task on the platform **P2** using novel elongated objects and geometrically diverse cups and mugs, including cases with partial occlusion. As shown in Fig. 13, the same anchoring rules remain effective across substantial shape and size variations and under dynamic perturbations. These results indicate that high-fidelity pose estimation is not essential for the considered manipulation relationships, and that simple object-centric anchors can provide an effective and scalable representation for cross-instance adaptation.

2) Interpolation Density: We further study the interpolation density n used during pre-grasping (refer Eqn. 7). Interpolated waypoints provide a smoother and safer approach, reduce premature end-effector-object contact, and create intermediate states for responding to external or execution-induced object offset. We utilized the dual-arm humanoid platform **P2** to conduct four bimanual tasks. As shown in Tab. VI, increasing n generally improves grasp stability, while the benefit saturates beyond a moderate density. This indicates that VLBiMan++ does not require excessive waypoint discretization. We use $n=6$ as a practical choice that balances pre-grasping reliability

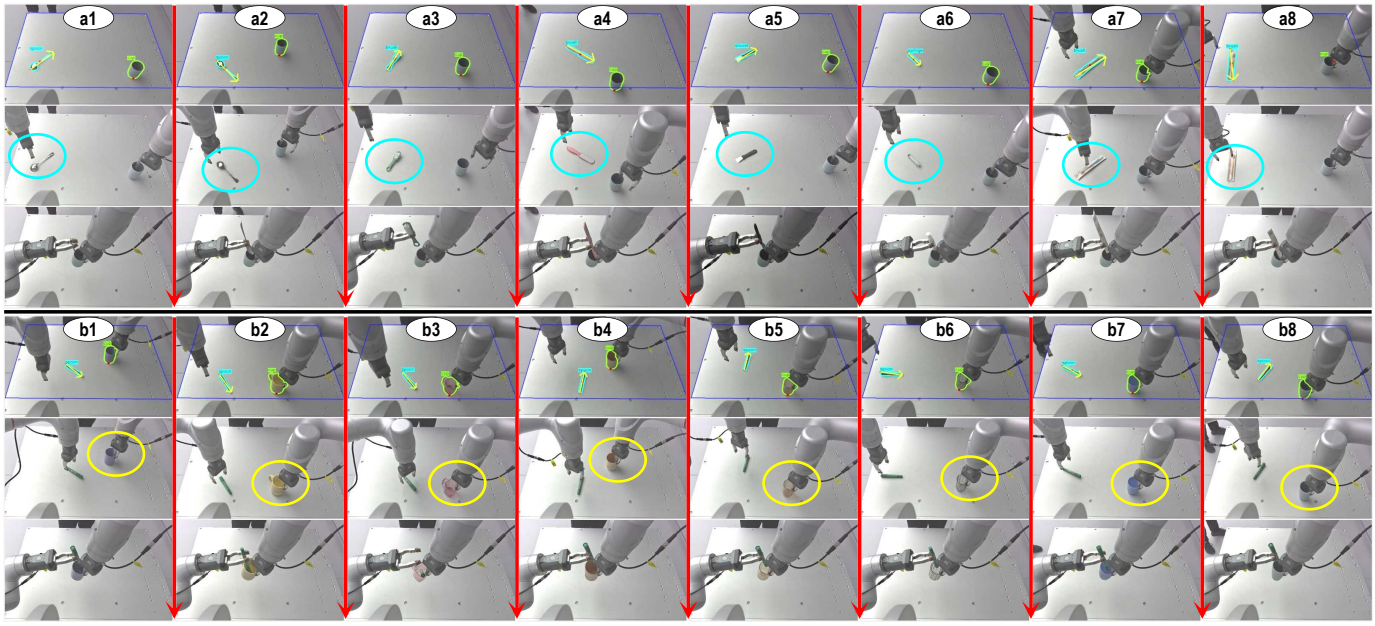


Fig. 13. Taking the `inserting` task as an example, we can replace the marker pen held by the left arm with other rectangular objects that were completely different (*e.g.*, *spoon* in **a1/a2**, *brush* in **a3/a4**, *spatula* in **a5**, *syringe* in **a6**, and *toothbrush* in **a7/a8**) as in the ***top row***, or change the cup grasped by the right arm to cups of different shapes (*e.g.*, *mugs with handles* in **b1/b2/b3/b4**, and *ordinary cups without handles* in **b5/b6/b7/b8**) as in the ***bottom row***.

TABLE VI
ABLATION EXPERIMENTS OF THE INTERPOLATION DENSITY n. TO ENSURE RELIABLE SEARCHING OF THE OPTIMAL n, WE DID NOT ADD ANY ADDITIONAL DYNAMIC INTERFERENCE IN EACH TRAIL.

Inter-polation Density n	*ID setting*					*OOD setting*				
	`inserting`	`unscrew`	`pouring`	`reorient`	Average Success Rate	`inserting`	`unscrew`	`pouring`	`reorient`	Average Success Rate
$n=3$	15/20	14/20	12/20	13/20	67.5%	13/20	12/20	11/20	11/20	58.8%
$n=4$	18/20	15/20	15/20	16/20	80.0%	17/20	14/20	14/20	14/20	73.8%
$n=5$	19/20	17/20	18/20	16/20	87.5%	18/20	16/20	17/20	15/20	82.5%
$n=6$	19/20	19/20	18/20	17/20	91.3%	19/20	18/20	17/20	16/20	**87.5%**
$n=7$	19/20	18/20	18/20	18/20	**92.5%**	18/20	19/20	17/20	16/20	**87.5%**
$n=8$	19/20	19/20	17/20	18/20	**92.5%**	19/20	18/20	16/20	17/20	**87.5%**

TABLE VII
ABLATION STUDIES OF VLBIMAN++. ALL TRIALS WERE COMPLETED ON SIX PRIMARY TASKS USING THE PLATFORM **P1** (**ID + INTERFERENCE**).

VLMs-based grounding	state-aware adaptation	IK-refinement	collision avoidance	closed-loop re-composition	Avg. SR
SAM+DINOv2	Ours	✓	✓	✓	35.8%
Ours	AnyGrasp	✓	✓	✓	31.7%
Ours	Ours	✗	✓	✓	29.2%
Ours	Ours	✓	✗	✓	34.2%
Ours	Ours	✓	✓	✗	40.8%
Ours	Ours	✓	✓	✓	**59.2%**

and execution efficiency across tasks.

3) Module Ablation: We first ablate major components of VLBiMan++ by replacing or removing VLM-based grounding, state-aware adaptation, trajectory refinement (IK-refinement + collision avoidance), or closed-loop re-composition while keeping the other pipeline unchanged to examine its contribution. As shown in Tab. VII, each component contributes consistently to overall performance. Replacing VLM grounding degrades semantic target localization, while replacing state-aware adaptation particularly with the popular non-semantic AnyGrasp [91] (where we find the one closest to the demo grasp pose from many proposals for fair comparison) affects novel object geometries and articulated or deformable configurations. Without refined trajectory refinement, reachability and collision failures increase, whereas removing closed-loop re-composition reduces robustness to object displacement during execution. These results verify that the proposed components are complementary and collectively support VLBiMan++'s generalization and robustness.

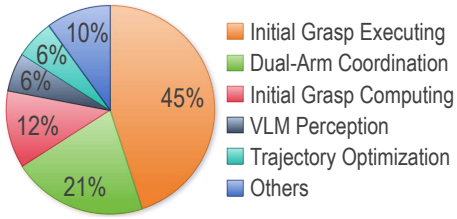


Fig. 14. Error breakdown performed on VLBiMan++ failures based on the rollout results in Tab. II (covering OOD and interference part).

4) Failure Analysis: We analyze remaining failures using four stages: Perception → *Planning*→ *Execution* → *Recovery*. The major failure categories consist of five compositions, consistent with the breakdown in Fig. 14. Among them, execution-related failures remain the dominant source, reflecting the difficulty of reliable physical grasping under kinematic singularities, imperfect contact, and early collisions. Dual-arm coordination constitutes another important source of failure, whereas VLM perception and trajectory optimization account for smaller proportions, supporting the reliability of the proposed grounding and trajectory synthesis modules. In the *Recovery* stage, the closed-loop mechanism can compensate for moderate object displacement and temporary perception uncertainty by re-observing and selectively re-instantiating affected skills, but it cannot yet recover from fundamentally unreachable configurations, severe grasp failures, or large unmodeled object-state changes. This analysis clarifies both the

practical strengths of VLBiMan++ and the failure boundaries that remain important for future improvement.

V. Conclusion and Limitation

In this work, we presented VLBiMan++, an extended framework for generalizable bimanual robotic manipulation that expands the reuse boundary of a single demonstration through vision-language anchoring, state-aware skill abstraction, and closed-loop trajectory composition. By decomposing demonstrations into state-invariant and state-conditioned skills, grounding task-relevant objects through vision-language perception, and selectively adapting their geometric or object-state-dependent components, VLBiMan++ avoids repeated data collection and policy retraining while supporting increasingly diverse conditions. Extensive real-world experiments demonstrate generalization across tasks and skill compositions, unseen rigid, articulated, and deformable objects, cluttered and dynamic scenes, heterogeneous dual-arm embodiments, and prolonged closed-loop execution. These results suggest that a carefully structured one-shot demonstration can serve not merely as a trajectory to imitate, but as a reusable task prior whose functional structure can be repeatedly re-instantiated across changing physical and environmental conditions.

Limitations: Despite these advances, several limitations remain. ***First***, the current object-state representation is intentionally compact and task-oriented, and therefore does not fully model highly complex or dynamically changing object configurations. Although VLBiMan++ extends beyond rigid objects to articulated and deformable cases, highly deformable materials with topology changes, self-occlusion, or strong dynamics may require richer representations and more frequent perception-action updates. ***Second***, the current closed-loop mechanism primarily detects observable state deviations and selectively re-adapts the affected skills. It does not yet provide a comprehensive failure-aware reasoning and recovery system capable of diagnosing whether an action has succeeded, identifying the cause of a failure, and autonomously selecting an alternative strategy. This remains particularly important for long-horizon manipulation, where small execution errors can accumulate across multiple contacts and skill transitions. ***Third***, collision handling is still largely structured around pre-grasp compensation and validated skill trajectories. More general collision-aware planning in densely cluttered and dynamically changing environments would benefit from dedicated model-based planners or online reactive controllers that explicitly consider the full configuration space of both arms and surrounding objects. ***Fourth***, the current task decomposition still involves lightweight human refinement when automatic keypose extraction is insufficient, which limits fully autonomous scaling to large collections of demonstrations. Reliable automatic segmentation of task- and state-dependent skills remains an open problem. ***Finally***, the capability of the system is bounded by the sensing and actuation of the underlying hardware. Fixed-base arms limit workspace mobility, parallel grippers restrict dexterous in-hand manipulation, and the absence of force or tactile sensing limits interaction with force-sensitive, fragile, or contact-rich objects. Future research should therefore explore richer multimodal sensing, dexterous and force-aware end-effectors, mobile or whole-body embodiments, autonomous skill discovery, and failure-aware closed-loop recovery, with the ultimate goal of moving from one-shot skill transfer toward persistent, self-correcting, and broadly capable robotic manipulation.

References

[1] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," in *2024 IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 653–660.

[2] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[3] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 6892–6903.

[4] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, X. He, X. Huang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. in 2025 ieee," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2025, pp. 3549–3556.

[5] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong *et al.*, "Openvla: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*, 2024.

[7] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," in *The Thirteenth International Conference on Learning Representations*, vol. 2025, 2025, pp. 54 877–54 910.

[8] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[9] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," in *8th Annual Conference on Robot Learning*, 2024.

[10] J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich *et al.*, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint arXiv:2405.02292*, 2024.

[11] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *8th Annual Conference on Robot Learning*, 2024.

[12] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1b: a diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[13] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[14] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[17] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829.

[18] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[19] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[20] J. Yang, Z. Cao, C. Deng, R. Antonova, S. Song, and J. Bohg, "Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning," in *8th Annual Conference on Robot Learning*, 2024.

[21] F. Xie, A. Chowdhury, M. De Paolis Kaluza, L. Zhao, L. Wong, and R. Yu, "Deep imitation learning for bimanual robotic manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2327–2337, 2020.

[22] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5150–5163, 2022.

[23] Z. Yuan, T. Wei, S. Cheng, G. Zhang, Y. Chen, and H. Xu, "Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning," in *8th Annual Conference on Robot Learning*, 2024.

[24] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in *8th Annual Conference on Robot Learning*, 2024.

[25] K. Fang, F. Liu, P. Abbeel, and S. Levine, "Moka: Open-world robotic manipulation through mark-based visual prompting," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[26] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction in robotics," in *8th Annual Conference on Robot Learning*, 2024.

[27] H. Zhou and K. Jia, "Vlbiman: Vision-language anchored one-shot demonstration enables generalizable bimanual robotic manipulation," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 112 330–112 360.

[28] L. P. Kaelbling and T. Lozano-Pérez, "Integrated task and motion planning in belief space," *The International Journal of Robotics Research*, vol. 32, no. 9-10, pp. 1194–1227, 2013.

[29] N. T. Dantam, Z. K. Kingston, S. Chaudhuri, and L. E. Kavraki, "An incremental constraint-based framework for task and motion planning," *The International Journal of Robotics Research*, vol. 37, no. 10, pp. 1134–1151, 2018.

[30] T. Migimatsu and J. Bohg, "Object-centric task and motion planning in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 844–851, 2020.

[31] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2022, pp. 13 081–13 088.

[32] G. Papagiannis, N. Di Palo, P. Vitiello, and E. Johns, "R+ x: Retrieval and execution from everyday human videos," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 8284–8290.

[33] J. Gao, X. Jin, F. Krebs, N. Jaquier, and T. Asfour, "Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks," in *2024 IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 16 850–16 857.

[34] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[35] J. Grannen, P. Sundaresan, B. Thananjeyan, J. Ichnowski, A. Balakrishna, V. Viswanath, M. Laskey, J. Gonzalez, and K. Goldberg, "Untangling dense knots by learning task-relevant keypoints," in *Conference on Robot Learning*. PMLR, 2021, pp. 782–800.

[36] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu, "Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 222–239.

[37] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao, "Rt-affordance: Affordances are versatile intermediate representations for robot manipulation," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 8249–8257.

[38] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," in *The Eleventh International Conference on Learning Representations*, 2023.

[39] Y. Tang, W. Huang, Y. Wang, C. Li, R. Yuan, R. Zhang, J. Wu, and L. Fei-Fei, "Uad: Unsupervised affordance distillation for generalization in robotic manipulation," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 3822–3831.

[40] A. Colomé and C. Torras, "Dimensionality reduction for dynamic movement primitives and application to bimanual manipulation of clothes," *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 602–615, 2018.

[41] T. Weng, S. M. Bajracharya, Y. Wang, K. Agrawal, and D. Held, "Fabricflownet: Bimanual cloth manipulation with a flow-based policy," in *Conference on Robot Learning*. PMLR, 2022, pp. 192–202.

[42] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," in *The Twelfth International Conference on Learning Representations*, vol. 2024, 2024, pp. 40 938–40 958.

[43] X. Zhang and A. Boularias, "One-shot imitation learning with invariance matching for robotic manipulation," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[44] T. Zhang, Y. Hu, H. Cui, H. Zhao, and Y. Gao, "A universal semantic-geometric representation for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 3342–3363.

[45] Y. Chen, C. Wang, Y. Yang, and K. Liu, "Object-centric dexterous manipulation from human motion data," in *8th Annual Conference on Robot Learning*, 2024.

[46] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *8th Annual Conference on Robot Learning*, 2024.

[47] J. Kerr, C. M. Kim, M. Wu, B. Yi, Q. Wang, K. Goldberg, and A. Kanazawa, "Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction," in *8th Annual Conference on Robot Learning*, 2024.

[48] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "Vividex: Learning vision-based dexterous manipulation from human videos," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 3336–3343.

[49] G. Ponimatkin, M. Cífka, T. Soucek, M. Fourmy, Y. Labbé, V. Petrik, and J. Sivic, "6d object pose tracking in internet videos for robotic manipulation," in *The Thirteenth International Conference on Learning Representations*, vol. 2025, 2025, pp. 32 569–32 596.

[50] W. Ye, F. Liu, Z. Ding, Y. Gao, O. Rybkin, and P. Abbeel, "Video2policy: Scaling up manipulation tasks in simulation through internet videos," *arXiv preprint arXiv:2502.09886*, 2025.

[51] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 306–324.

[52] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu, "Oakink2: A dataset of bimanual hands-object manipulation in complex task completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 445–456.

[53] Y. Liu, H. Yang, X. Si, L. Liu, Z. Li, Y. Zhang, Y. Liu, and L. Yi, "Taco: Benchmarking generalizable bimanual tool-action-object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 740–21 751.

[54] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 383–19 400.

[55] H. Zhao, X. Liu, M. Xu, Y. Hao, W. Chen, and X. Han, "Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2025, pp. 27 683–27 693.

[56] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 13 226–13 233.

[57] B. Wen, W. Lian, K. Bekris, and S. Schaal, "You only demonstrate once: Category-level manipulation from single visual demonstration," in *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022.

[58] A. Bahety, P. Mandikal, B. Abbatematteo, and R. Martín-Martín, "Screwmimic: Bimanual imitation from human videos with screw space projection," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[59] H. Zhou, R. Wang, Y. Tai, Y. Deng, G. Liu, and K. Jia, "You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[60] Y. Wang and E. Johns, "One-shot dual-arm imitation learning," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 5660–5668.

[61] J. Mao, T. Lozano-Pérez, J. B. Tenenbaum, and L. P. Kaelbling, "Learning reusable manipulation strategies," in *Conference on Robot Learning*. PMLR, 2023, pp. 1467–1483.

[62] Y. Liu, J. Mao, J. B. Tenenbaum, T. Lozano-Pérez, and L. P. Kaelbling, "One-shot manipulation strategy learning by making contact analogies," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 15 387–15 393.

[63] O. Biza, S. Thompson, K. R. Pagidi, A. Kumar, E. van der Pol, R. Walters, T. Kipf, J.-W. van de Meent, L. L. Wong, and R. Platt, "One-shot imitation learning via interaction warping," in *Conference on Robot Learning*. PMLR, 2023, pp. 2519–2536.

[64] H. Zhou and K. Jia, "One-shot real-world demonstration synthesis for scalable bimanual manipulation," in *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026.

[65] H. Zhou, R. Wang, Y. Tai, Y. Deng, G. Liu, and K. Jia, "Yoto++: Learning long-horizon closed-loop bimanual manipulation from one-shot human video demonstrations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 9, pp. 10 695–10 712, 2026.

[66] Y. J. Ma, J. Hejna, C. Fu, D. Shah, J. Liang, Z. Xu, S. Kirmani, P. Xu, D. Driess, T. Xiao *et al.*, "Vision language models are in-context value learners," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 33 984–34 009.

[67] H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, and Z. Zhao, "Roboground: Robotic manipulation with grounded vision-language priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 540–22 550.

[68] H. Fang, M. Grotz, W. Pumacay, Y. R. Wang, D. Fox, R. Krishna, and J. Duan, "Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation," in *Forty-second International Conference on Machine Learning*, 2025.

[69] Y. Feng, J. Han, Z. Yang, X. Yue, S. Levine, and J. Luo, "Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation," in *Conference on Robot Learning*. PMLR, 2025, pp. 2038–2062.

[70] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation*. IEEE, 2023, pp. 9493–9500.

[71] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation*. IEEE, 2023, pp. 11 523–11 530.

[72] A. Szot, M. Schwarzer, H. Agrawal, B. Mazoure, R. Metcalf, W. Talbott, N. Mackraz, R. D. Hjelm, and A. T. Toshev, "Large language models as generalizable policies for embodied tasks," in *The Twelfth International Conference on Learning Representations*, 2024.

[73] H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao, "Copa: General robotic manipulation through spatial constraints of parts with foundation models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2024, pp. 9488–9495.

[74] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

[75] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *The Thirteenth International Conference on Learning Representations*, vol. 2025, 2025, pp. 28 085–28 128.

[76] Y. Kuang, J. Ye, H. Geng, J. Mao, C. Deng, L. Guibas, H. Wang, and Y. Wang, "Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation," in *8th Annual Conference on Robot Learning*, 2024.

[77] K. Dreczkowski, P. Vitiello, V. Vosylius, and E. Johns, "Learning a thousand tasks in a day," *Science Robotics*, vol. 10, no. 108, p. eadv7594, 2025.

[78] A. Patel, B. Pekarek, J. E. C. Hernandez, and S. Song, "Behavior prompting policy: Demonstrations as prompts for manipulation," *arXiv preprint arXiv:2606.30457*, 2026.

[79] G. Chen, M. Wang, T. Cui, Z. Zhou, Q. Shao, S. Li, H. Su, R. Gan, H. Wang, M. Fu *et al.*, "Host: Robots acquire manipulation skills in seconds from a single human video," *arXiv preprint arXiv:2607.20033*, 2026.

[80] S. James, K. Wada, T. Laidlow, and A. J. Davison, "Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 739–13 748.

[81] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[82] X. Ma, S. Patidar, I. Haughton, and S. James, "Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 081–18 090.

[83] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in *Conference on Robot Learning*, 2024.

[84] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[85] J. Schulman, Y. Duan, J. Ho, A. Lee, I. Awwal, H. Bradlow, J. Pan, S. Patil, K. Goldberg, and P. Abbeel, "Motion planning with sequential convex optimization and convex collision checking," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1251–1270, 2014.

[86] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.

[87] G. Xu, X. Wang, Z. Zhang, J. Cheng, C. Liao, and X. Yang, "Igev++: Iterative multi-range geometry encoding volumes for stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 7108–7122, 2025.

[88] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 906–27 916.

[89] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.

[90] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 444–11 453.

[91] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[92] F. Chaumette, "Image moments: a general and useful set of features for visual servoing," *IEEE Transactions on Robotics*, vol. 20, no. 4, pp. 713–723, 2004.

[93] L. Kotoulas and I. Andreadis, "Accurate calculation of image moments," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2028–2037, 2007.

[94] H. Zhou, F. Jiang, and H. Lu, "Body-part joint detection and association via extended object representation," in *2023 IEEE International Conference on Multimedia and Expo*. IEEE, 2023, pp. 168–173.

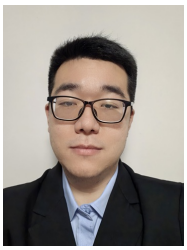
[95] H. Zhou, F. Jiang, J. Si, Y. Ding, and H. Lu, "Bpjdet: Extended object representation for generic body-part joint detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4314–4330, 2024.

[96] G. Bradski, "The opencv library." *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[97] E. W. Dijkstra, "A note on two problems in connexion with graphs," in *Edsger Wybe Dijkstra: his life, work, and legacy*, 2022, pp. 287–290.

[98] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

Huayi Zhou is now an Assistant Professor at the College of Computer Science and Software Engineering, Shenzhen University. Before that, he was a postdoctoral researcher at The Chinese University of Hong Kong, Shenzhen during years 2024-2026. He received his B.S degree at Dept. of Computer Science from Hunan University in 2017, and both M.S. degree and Ph.D. degree at Dept. of Computer Science and Engineering from Shanghai Jiao Tong University in 2020 and 2024, respectively. His research interests lie in emboided robot manipulation, multi-modal perception, and computer vision (e.g., object detection, pose estimation, and monocular 3D reconstruction), combined with enduring machine learning techniques such as multi-task learning, domain adaptation, domain generalization and semi-supervised learning.



Wei Gao is now the Chief Scientist at Dexforce Robotics. Before that, he was the Chief Scientist at Mech-Mind Robotics during years 2021-2024. He received his B.S. degree at the School of Aerospace Engineering from Tsinghua University in 2017, and his Ph.D. degree at the Department of Electrical Engineering and Computer Science from Massachusetts Institute of Technology (MIT) in 2021. His research interests lie in robotics, with sub-fields include manipulation, 3D vision, motion planning & control, and robotic foundation models (such as Vision-Language-Action (VLA) and World-Action Model (WAM)).

Yiyang Han is currently an undergraduate student at the Faculty of Engineering, Department of Computing, Imperial College London. His research interests are in machine learning and artificial intelligence. His recent research focuses on real-time turn-taking in spoken dialogue systems, parameter-efficient imitation learning for human behavior modeling, and embodied robot perception and manipulation.

Kui Jia is currently a Professor with School of Data Science, The Chinese University of Hong Kong, Shenzhen. He received the B.Eng. degree in marine engineering from Northwestern Polytechnical University, China, in 2001, the M.Eng. degree in electrical and computer engineering from the National University of Singapore in 2003, and the Ph.D. degree in computer science from Queen Mary University of London, London, U.K., in 2007. His research interests are in computer vision and machine learning. His recent research focuses on theoretical deep learning and its applications to 3D vision. He has been serving as an Associate Editor for IEEE Trans. on Image Processing and Trans. on Machine Learning Research.

Hui Huang is a Chair Professor at Shenzhen University, serving as the Dean of CS while also directing the Visual Computing Research Center (VCC). Her research spans computer graphics, computer vision, and visual analytics, with a particular focus on geometry. She has held editorial board positions for ACM TOG and IEEE TVCG. She is an inducted member of the ACM SIGGRAPH Academy and a CSIG Fellow. She has also been appointed Technical Papers Chair for ACM SIGGRAPH Asia 2027. Her homepage is: https://vcc.tech/huihuang.